Formation à l'analyse statistique

Module 1 : Statistique élémentaire

- **A** Indicateurs statistiques
- **B** Représentations graphiques
- **C** Notions de probabilités
- **D** Estimation et test statistique
- **E** Modélisation

Sébastien Déjean - math.univ-toulouse.fr/~sdejean

A Indicateurs statistiques

- Indicateurs pour 1 série de données
  - Position / Tendance centrale
  - Dispersion
- Indicateurs pour 2 séries de données

Objectif

Résumer une série de chiffres par un ou plusieurs chiffres (mais pas beaucoup)

Sébastien Déjean - math.univ-toulouse.fr/~sdejean

Indicateurs de tendance centrale

Moyenne : somme des observations divisées par le nombre d'observations $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Médiane : Valeur qui sépare l'échantillon en 2 sous-ensembles de tailles égales.

Mode : Valeur la plus fréquente dans un échantillon.

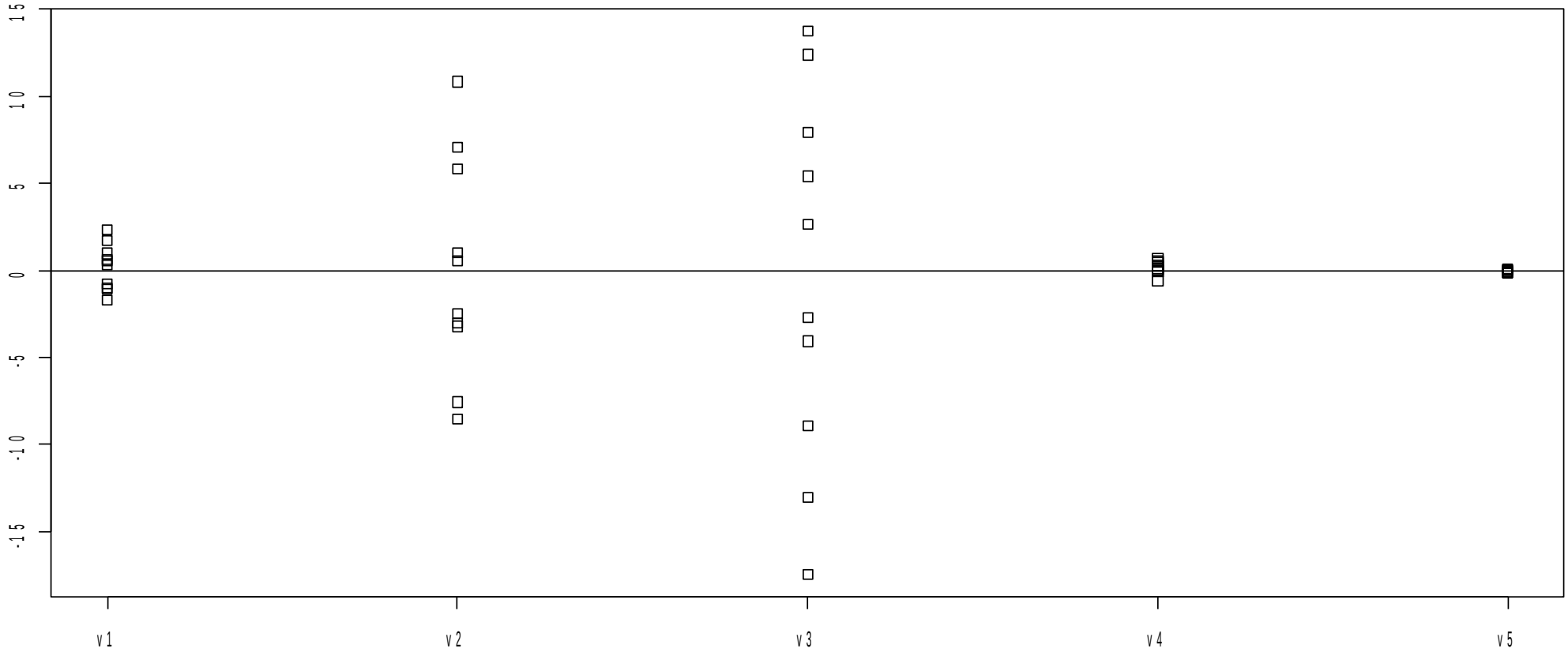
Quartiles : **3** valeurs qui sépare l'échantillon en **4** sous-ensembles de tailles égales.

Déciles : **9** valeurs qui sépare l'échantillon en **10** sous-ensembles...

Percentiles : **99** valeurs qui sépare l'échantillon en **100** sous-ensembles...

Quantiles : généralise les précédents

Limites des indicateurs de position



Les 5 séries de 10 observations représentées ici ont, à peu près, les mêmes moyennes et médianes mais on voit bien qu'elles ne se « ressemblent » pas. Les valeurs de chaque série sont plus ou moins dispersées.

Indicateurs de dispersion

Variance

$$\mathrm{var}(X)=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$$

Écart-type

$$\sigma(X)=\sqrt{\mathrm{var}(X)}$$

.....

Étendue : différence entre la plus grande et la plus petite valeur d'un échantillon

Espace inter-quartile : différence entre le 1er et le 3ème quartile,
correspond à l'étendue de l'échantillon privé de la moitié de ces observations (le ¼ le plus élevé et le ¼ le plus faible)

Variance et écart-type

$$\mathrm{var}(X)=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$$ Moyenne des carrés des écarts à la moyenne

$$\sigma(X)=\sqrt{\mathrm{var}(X)}$$ Racine carrée de la variance

Quelques propriétés de l'écart-type :

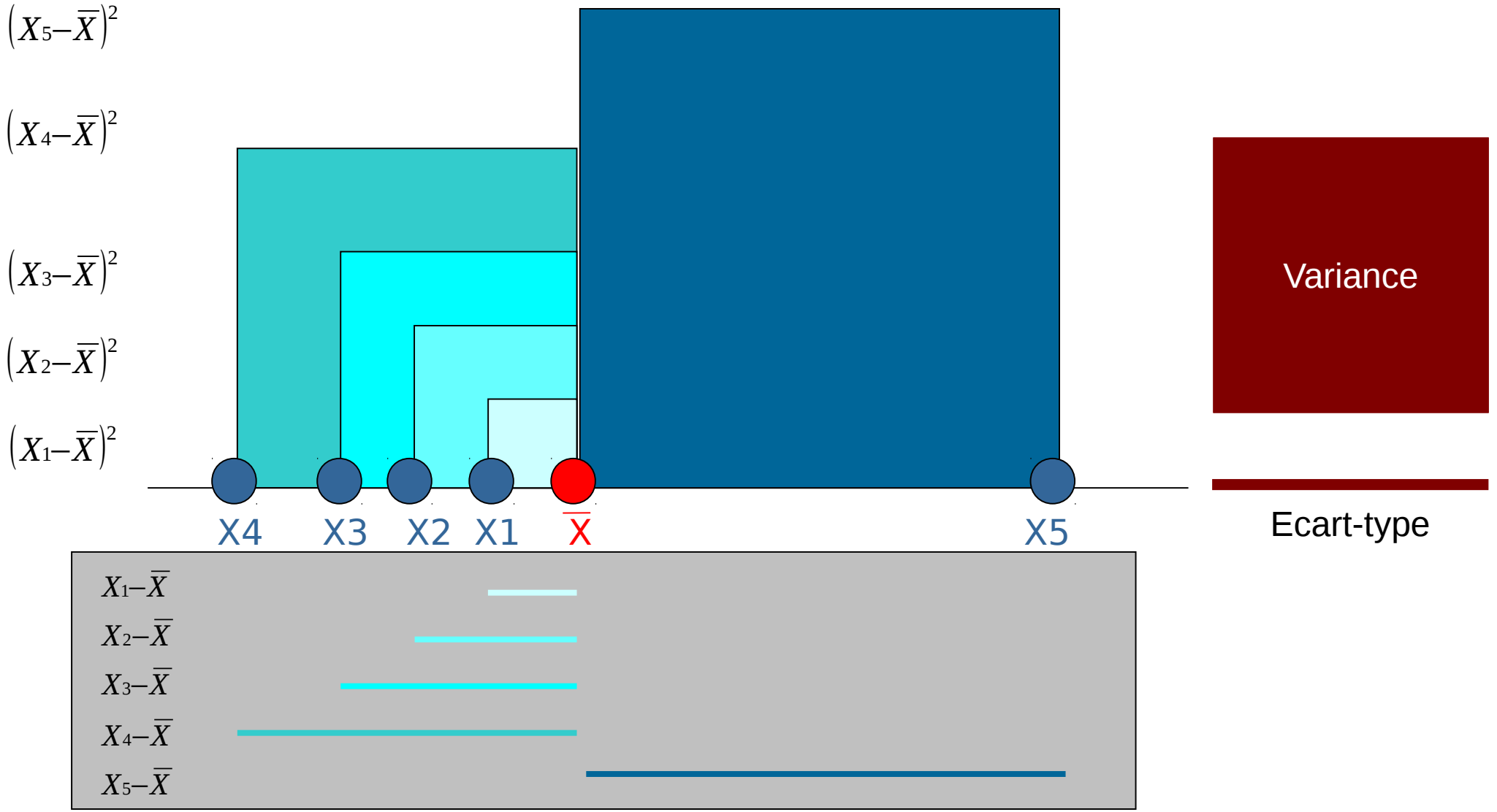
- Positif (nul si la série est constante)
- Invariant par translation
- Sensible aux valeurs extrêmes
- **De la même unité que la donnée** (et que la moyenne) :

*Si l'échantillon est constitué de mesures en **m** alors l'écart-type s'exprime également en **m** (tout comme la moyenne) ; ce qui n'est pas le cas de la variance **m²** !*

On peut ainsi additionner moyenne et écart-type (*mais pas moyenne et variance*), ce qui est fondamental pour la construction d'intervalle de confiance.

Variance et écart-type

Racine carrée de la moyenne des carrés des écarts à la moyenne



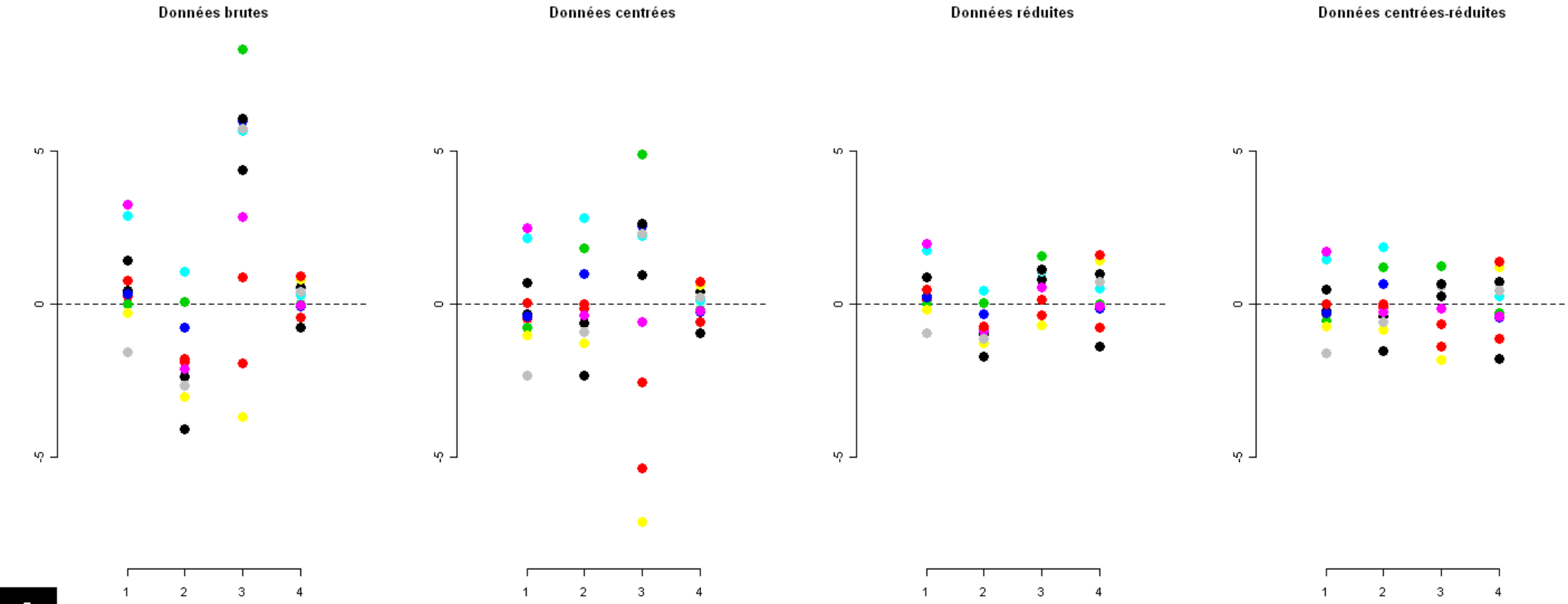
Sébastien Déjean - math.univ-toulouse.fr/~sdejean

A

Centrage-Réduction

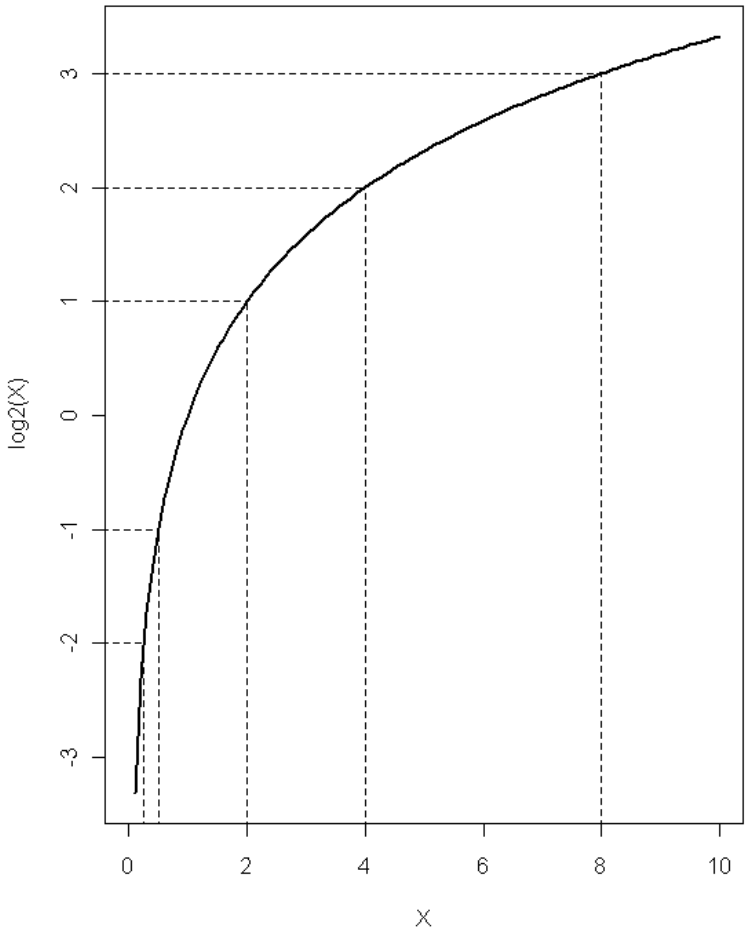
$$Z_i = \frac{X_i - \overline{X}}{\sigma_X}$$

- **Centrer** : retrancher la moyenne
- **Réduire** : diviser par l'écart-type
- Permet d'exprimer des variables différentes sur une échelle commune, en les débarrassant de leurs unités physiques : les observations s'expriment en **nombre d'écart-type par rapport à la moyenne**.
- Après centrage-réduction, la moyenne des observations est nulle et l'écart-type vaut 1 (ainsi que la variance).
- Appelé parfois « z-transformation » ou « z-score »



Conversion « log »

X	0,125 $= 2^{-3}$	0,25 $= 2^{-2}$	0,5 $= 2^{-1}$	1 $= 2^{0}$	2 $= 2^{1}$	4 $= 2^{2}$	8 $= 2^{3}$
$\log_2(X)$	-3	-2	-1	0	1	2	3



- Utile pour la conversion de ratio
- Exemple : Une sur-expression double et une sous-expression de moitié sont rendues symétriques par rapport à 0
- Pour les p-values, on aura plus intérêt à utiliser $\log_{10}$.
- La fonction réciproque de « log » est la fonction puissance :

$$Y = \log_2(X) \leftrightarrow X = 2^Y$$

$$Y = \log_{10}(X) \leftrightarrow X = 10^Y$$

$$Y = \ln(X) \leftrightarrow X = e^Y = \exp(Y)$$

Indicateurs statistiques 2D

Covariance $$\mathrm{cov}(X,Y)=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)$$

cov(X,X)=var(X)

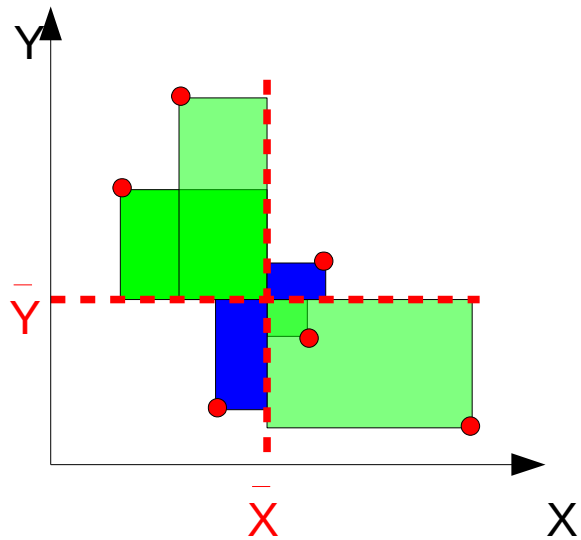
Signe du produit (Xi-$\overline{X}$)(Yi-$\overline{Y}$)

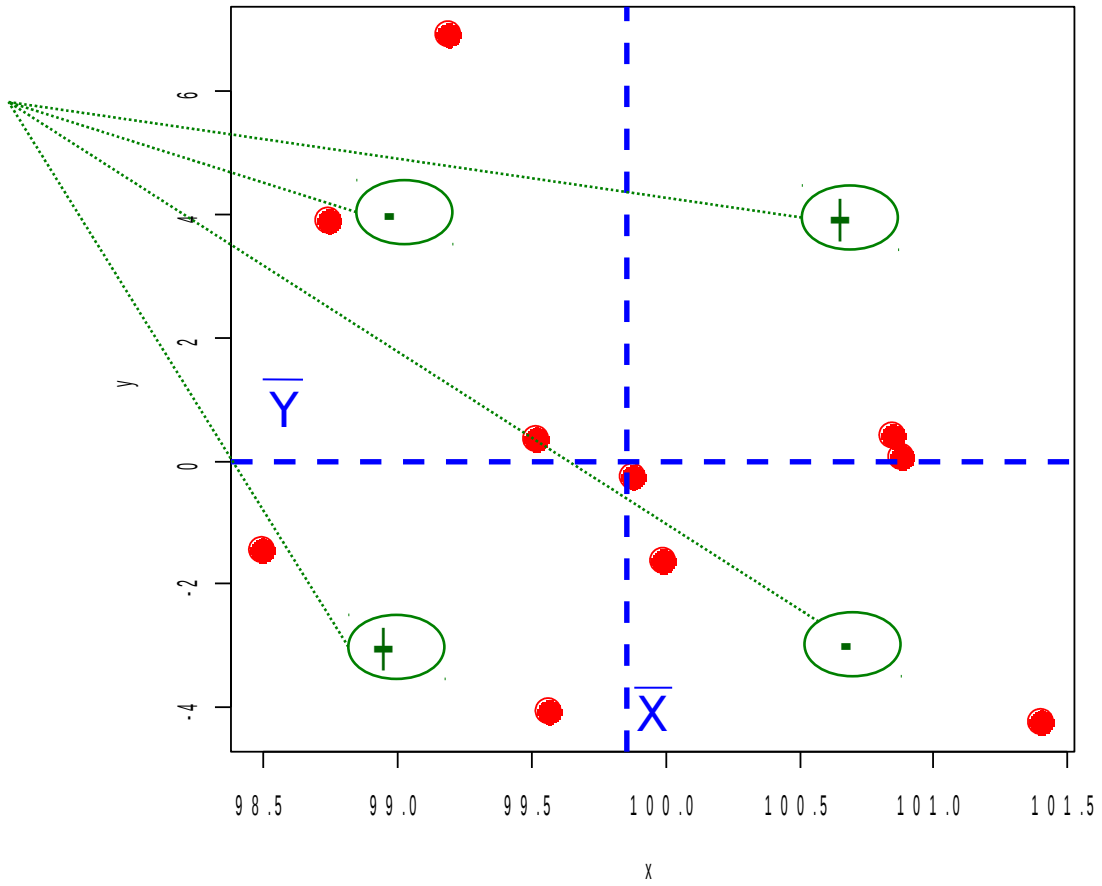
Intuitivement :

- Si les **+** l'emportent
  → liaison linéaire positive

- Si les **–** l'emportent
  → liaison linéaire négative

*Sur cet exemple : cov(X,Y)=**-1.36***





La covariance dépend des unités de mesure ➔ coefficient de corrélation

Coefficient(s) de corrélation

- Coefficient de corrélation **linéaire** de Pearson

$$\rho(X,Y) = cov(X,Y)/(\sigma_X \sigma_Y)$$

- Coefficient de corrélation de Spearman
  ➔ *Robustesse due au travail sur les rangs*

$$\rho_s(X,Y) = \rho(RX,RY) = 1 - 6\frac{\sum_{i=1}^{n} d^2}{n(n^2-1)}$$

(1) *calcul des rangs*

(2) *différence des rangs*

(3) *carrés des différences des rangs*

(4) *somme des carré des différences des rangs*

			(1)		**(2)**	**(3)**
X	Y		RX	RY	d = RX-RY	d²
20,6	20,7		6	7	-1	1
21,6	21,8		2	2	0	0
18,8	20,4		9	8	1	1
20,8	21,1		3	4	-1	1
17,5	18,3		10	10	0	0
19,5	18,9		7	9	-2	4
20,8	21,1		4	4	0	0
20,6	21,2		5	3	2	4
19,2	20,9		8	6	2	4
22,2	22,9		1	1	0	0
					Somme	**15**

(4)

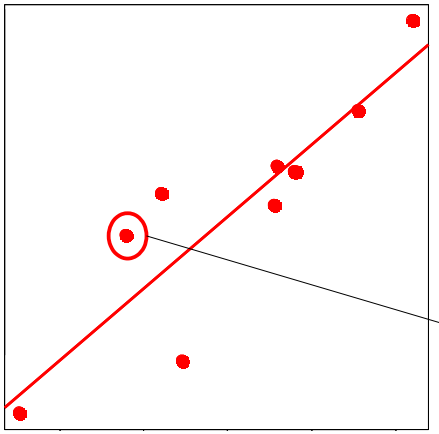
Coefficient(s) de corrélation

Quelques propriétés des coefficients de corrélation :

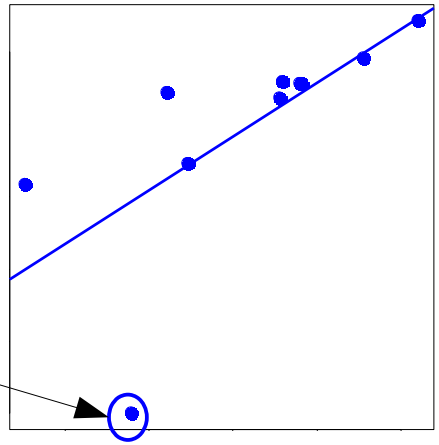
- Coefficient de corrélation de Pearson : **relation linéaire**
- Coefficient de corrélation de Spearman : considère les rangs, relation "**monotone**"
- Compris entre –1 et 1.
- Les valeurs extrêmes –1 et 1 indique des corrélations parfaites entre les 2 variables.
- Si le coefficient est **positif** : quand une variable est élevée, l'autre l'est également. Quand une variable est faible, l'autre l'est également.
- Si le coefficient est **négatif** : quand une variable est élevée (resp. faible), l'autre est faible (resp. élevée).

Corrélation : exemples

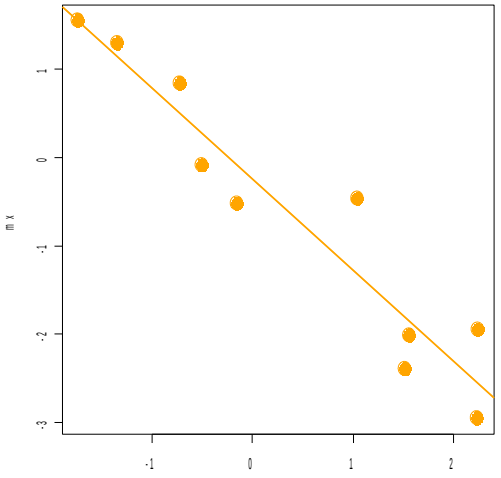
ρ : Pearson - ρ_s : Spearman



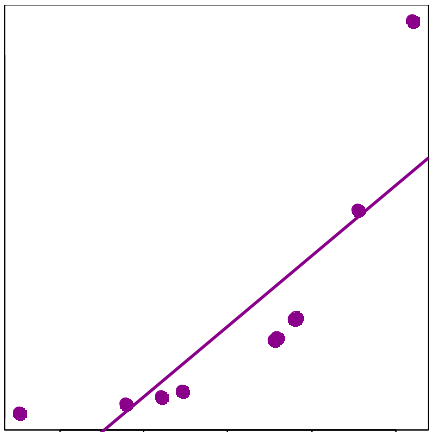
ρ=0.884 - ρ_s=0.9



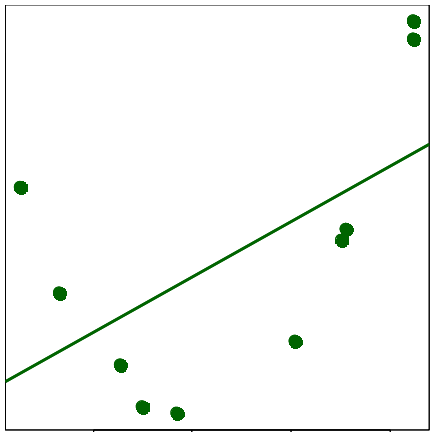
ρ=0.676 - ρ_s=0.912



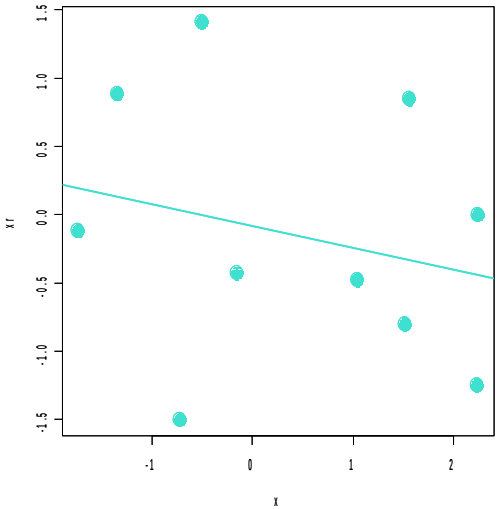
ρ=-0.954 - ρ_s=-0.903



ρ = 0.822 - ρ_s=1



ρ=0.584 - ρ_s=0.491

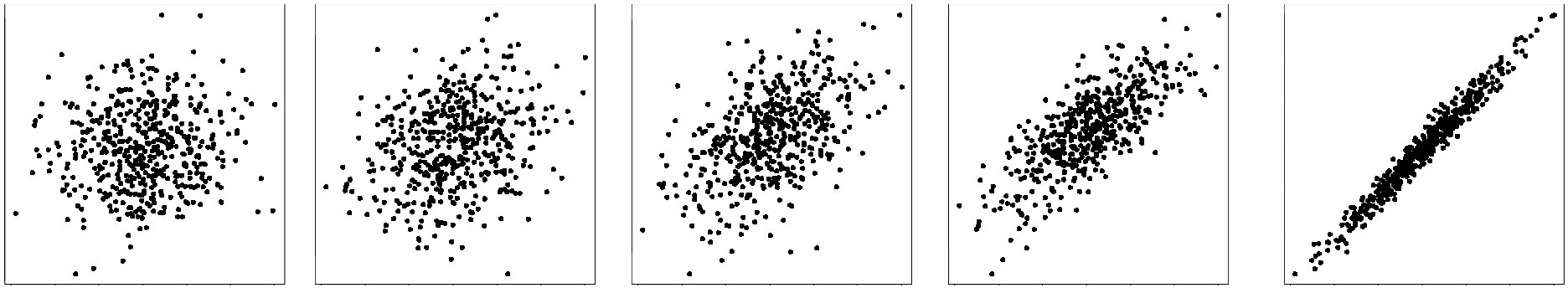


ρ=-0.248 - ρ_s=-0.164

Sébastien Déjean - math.univ-toulouse.fr/~sdejean

Corrélation : exemples

0.05 0.26 0.49 0.75 0.98



Valeurs significatives du coefficient de corrélation

v / α	0.10	0.05	0.02	v / α	0.10	0.05	0.02
1	0.9877	0.9969	0.9995	**16**	0.4000	0.4683	0.5425
2	0.9000	0.9500	0.980	**17**	0.3887	0.4555	0.5285
3	0.8054	0.8783	0.9343	**18**	0.3783	0.4438	0.5155
4	0.7293	0.8114	0.8822	**19**	0.3687	0.4329	0.5034
5	0.6694	0.7545	0.8329	**20**	0.3598	0.4227	0.4921
6	0.6215	0.7067	0.7887	**25**	0.3233	0.3809	0.4451
7	0.5822	0.6664	0.7498	**30**	0.2960	0.3494	0.4093
8	0.5494	0.6319	0.7155	**35**	0.2746	0.3246	0.3810
9	0.5214	0.6021	0.6851	**40**	0.2573	0.3044	0.3578
10	0.4973	0.5750	0.6581	**45**	0.2428	0.2875	0.3384
11	0.4762	0.5529	0.6339	**50**	0.2306	0.2732	0.3218
12	0.4575	0.5324	0.6120	**60**	0.2108	0.2500	0.2948
13	0.4409	0.5139	0.5923	**70**	0.1954	0.2319	0.2737
14	0.4259	0.4973	0.5742	**80**	0.1829	0.2172	0.2565
15	0.4124	0.4821	0.5577	**90**	0.1726	0.2050	0.2422
				100	0.1638	0.1946	0.2301

Corrélation ≠ causalité

Quelques exemples d'événements corrélés n'impliquant pas forcément une causalité :

- Le fait de dormir avec ses chaussures est fortement corrélé avec le fait de se réveiller avec la « gueule de bois ». Doit-on en conclure que dormir avec ses chaussures donne la « gueule de bois » ? Ou un troisième facteur est-il impliqué ?
- La fréquence des attaques de requins est fortement corrélée à la vente de glaces sur les pages ! Manger des glaces rend-il plus appétissant pour les requins ?
- Les personnes qui meurent ont très fréquemment vu un médecin dans les jours qui ont précédé. Est-il si dangereux de rencontrer un médecin ?
- Dans la plupart des villes, on constate une forte corrélation positive entre le nombre de nids de cigognes et la natalité. Nous cacherait-on des choses ?
- …

Un peu de théorie

Un graphique devrait :

- Montrer les données
- Inciter celui qui regarde à penser
- Éviter de distordre ce que les donnés ont à dire
- Présenter beaucoup de données sur une petite surface
- Encourager l'œil à comparer différents morceaux des données
- Révéler les données à des niveaux différents : d'un aperçu global à des structures plus fines
- Servir un objectif clair et raisonnable
- Être étroitement intégré à une description statistique du jeu de données

Edward Tufte, *The Visual Display of Quantitative Information*, Cheshire, CT, Graphics Press, 2001, 2e éd. (1re éd. 1983)

Représentations graphiques

Données de type « effectif »

A	30
B	15
C	30
D	20
E	25

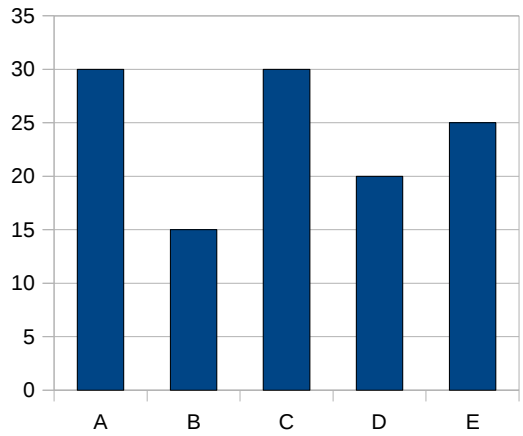
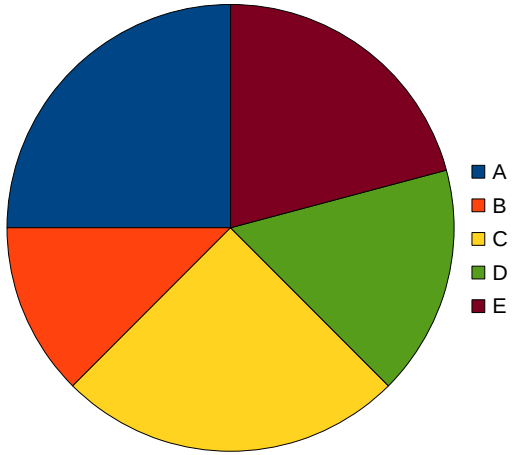
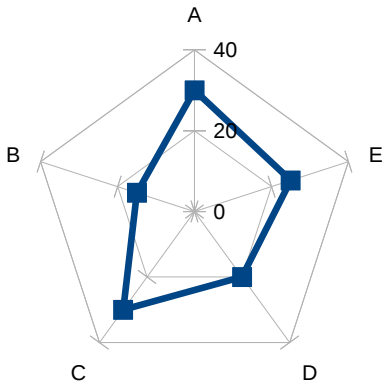


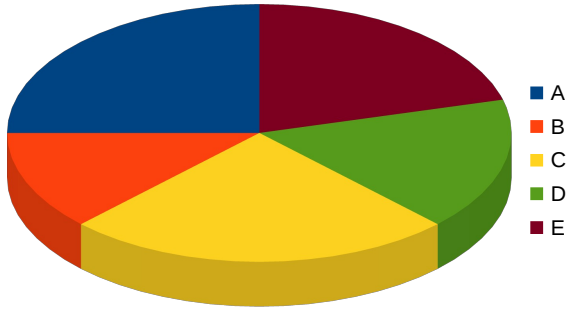
Diagramme en bâtons

Diagramme circulaire



Spiderman plot :-)

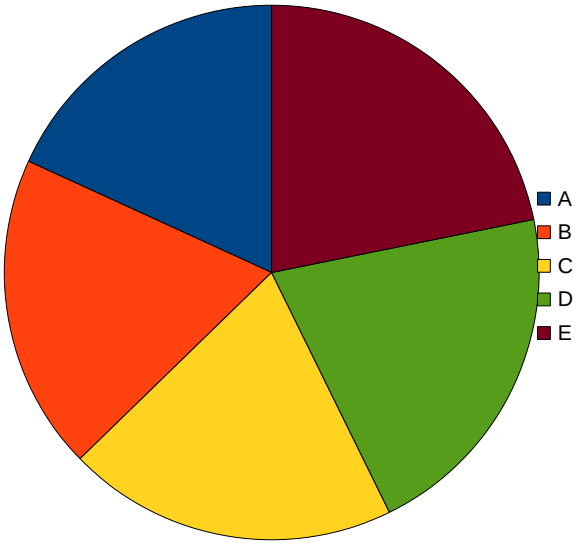




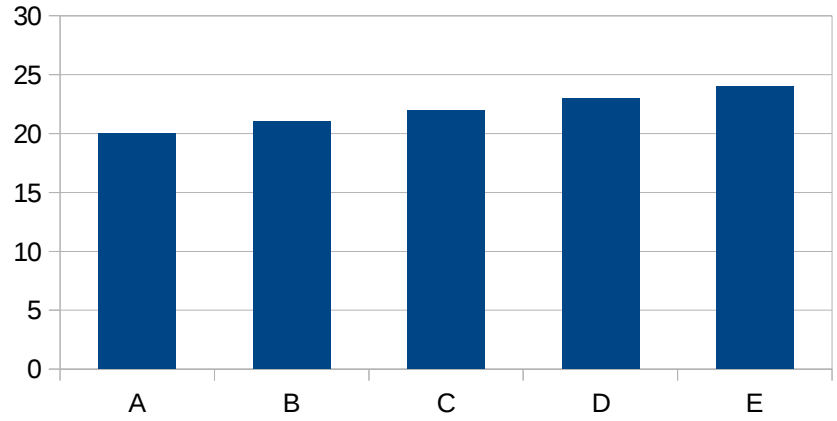
Attention aux représentations 3D. Est-ce si évident que A et C ont la même valeur ?

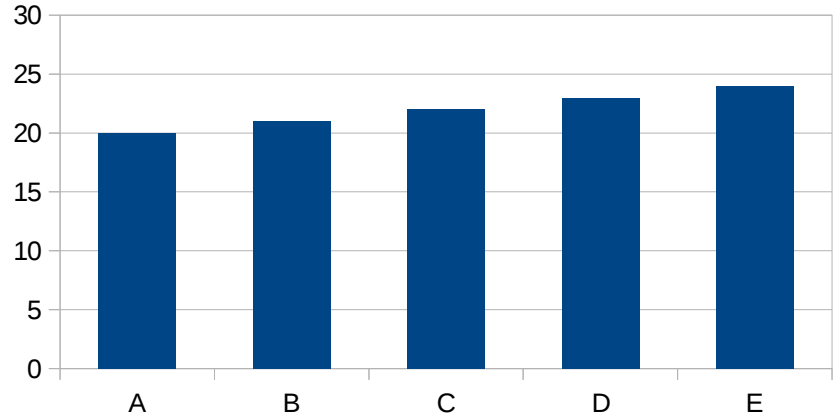
Représentations graphiques

Données de type « effectif »



A	20
B	21
C	22
D	23
E	24





B

Représentations graphiques

Données de type « effectif »

	X	Y
A	30	25
B	15	15
C	30	20
D	20	30
E	25	35

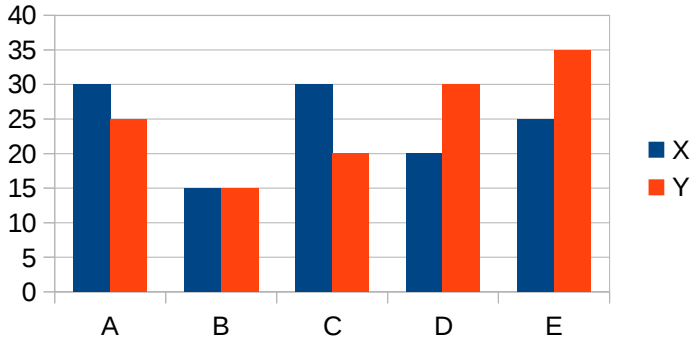
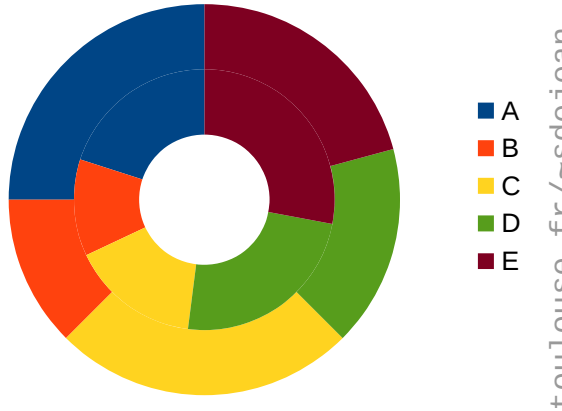
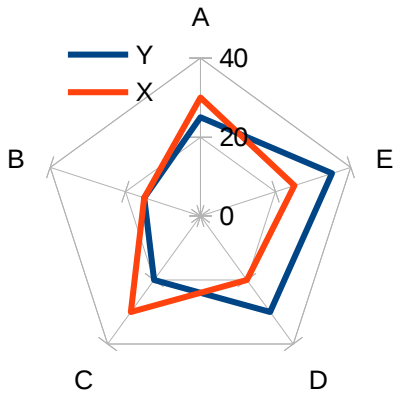


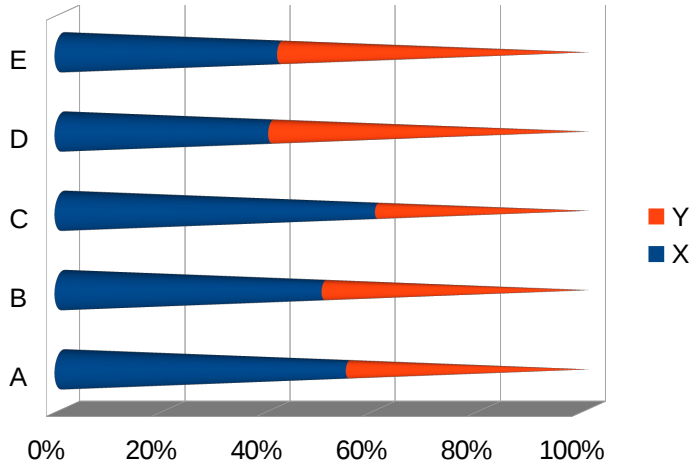
Diagramme en bâtons

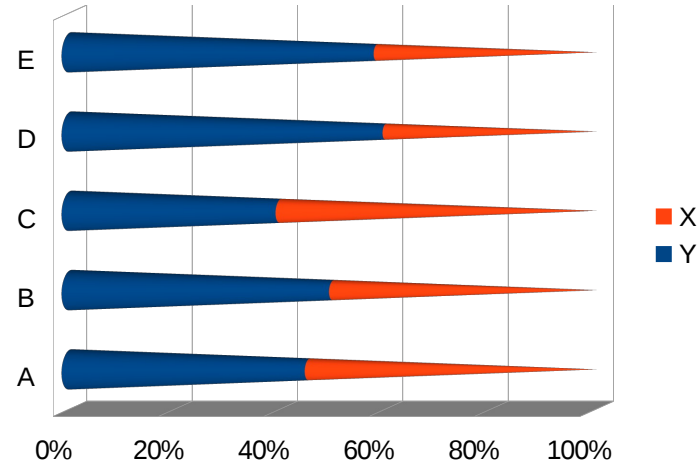
Diagramme « tore »





Cylindres 3D (!?)





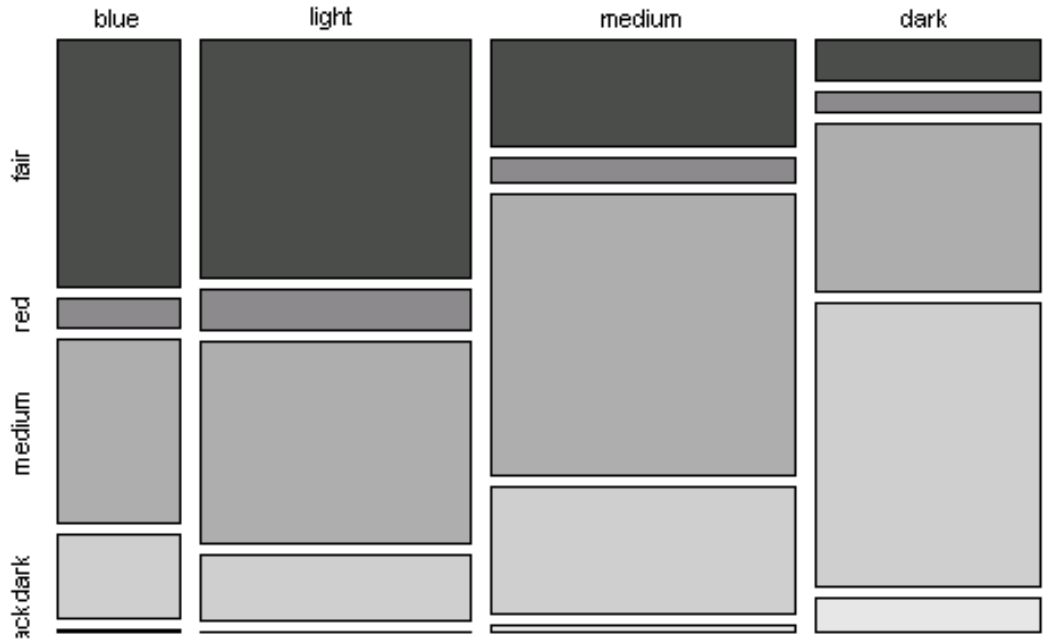
Représentations graphiques

Données de type « effectif » - Table de contingence

(données *caith*, R)

	fair	red	medium	dark	black
blue	326	38	241	110	3
light	688	116	584	188	4
medium	343	84	909	412	26
dark	98	48	403	681	85

Diagramme mosaïque



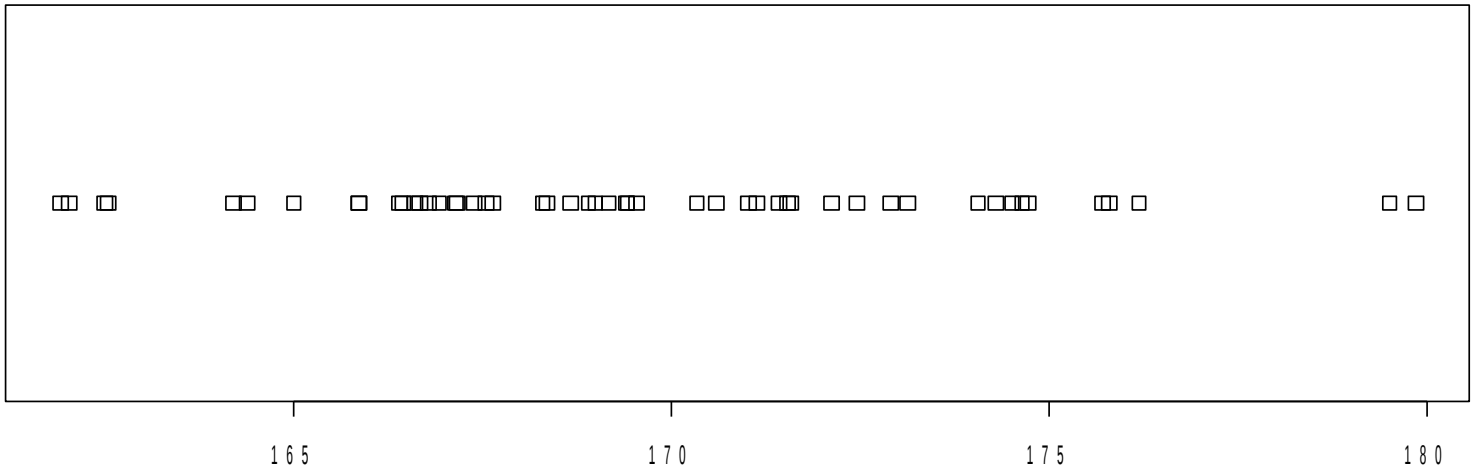
Représentations graphiques

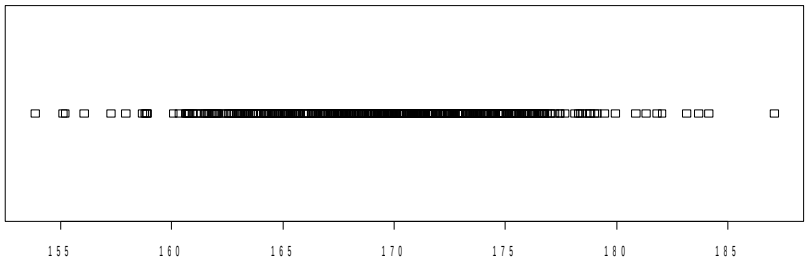
Données quantitatives

50 observations : taille en cm de 50 individus

171.4 175.8 168.9 166.8 174.3 169.2 165.9 171.6 167.1 179.8
174.5 169.0 170.6 173.1 169.4 167.4 162.5 172.9 174.7 164.2
175.7 167.2 169.5 174.1 162.5 172.4 166.4 162.0 167.5 168.7
166.9 171.0 176.2 172.1 166.4 168.3 170.3 164.4 167.6 168.3
165.9 174.6 171.5 169.4 166.6 179.5 166.7 171.1 161.9 165.0

Nuage de points 1D
(*stripchart*)

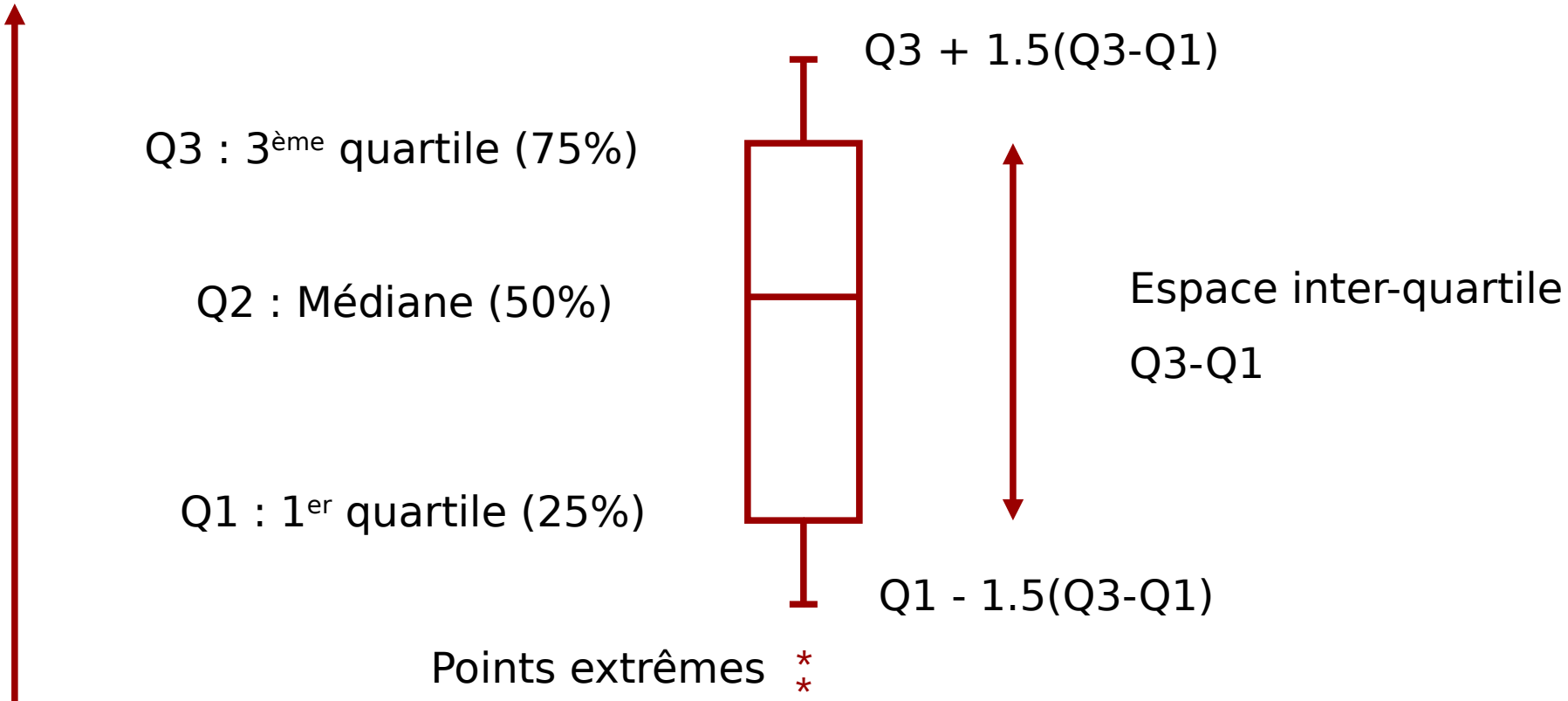


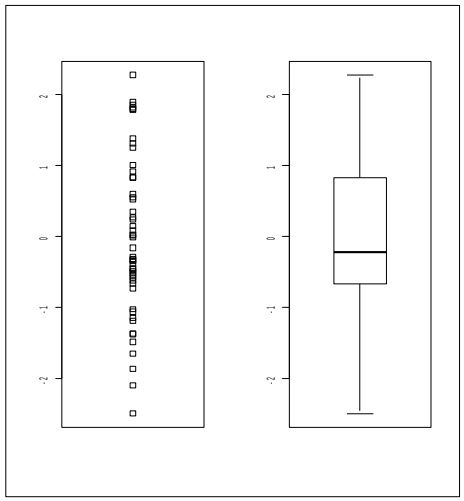


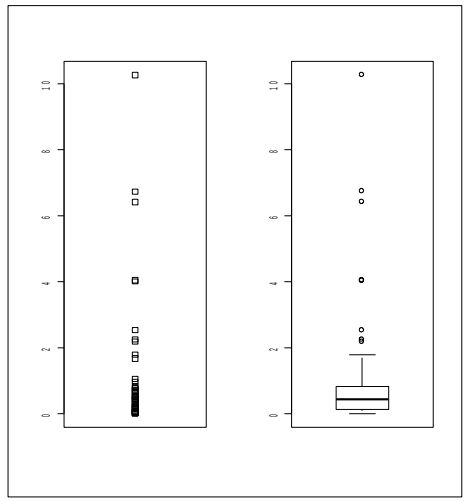
Avec 500 observations, la lisibilité devient délicate.

B

Boxplot



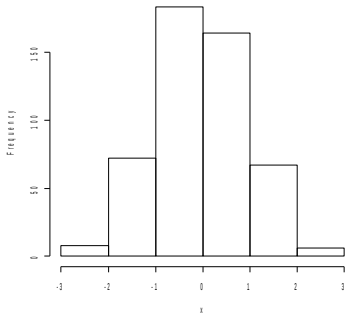


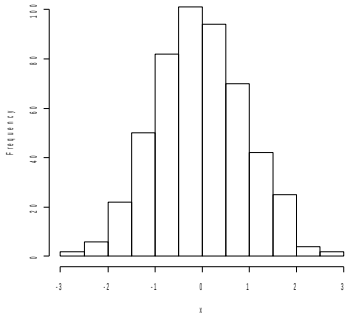


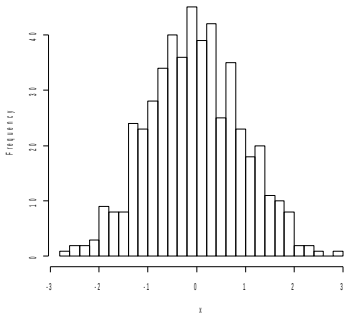
Histogramme

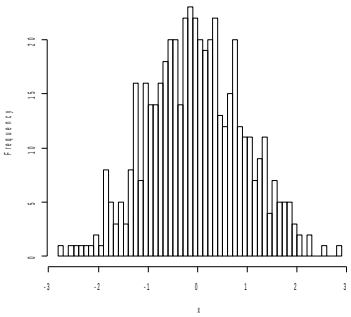
≠

diagramme en bâtons ☠









Histogramme

~

Densité de probabilité

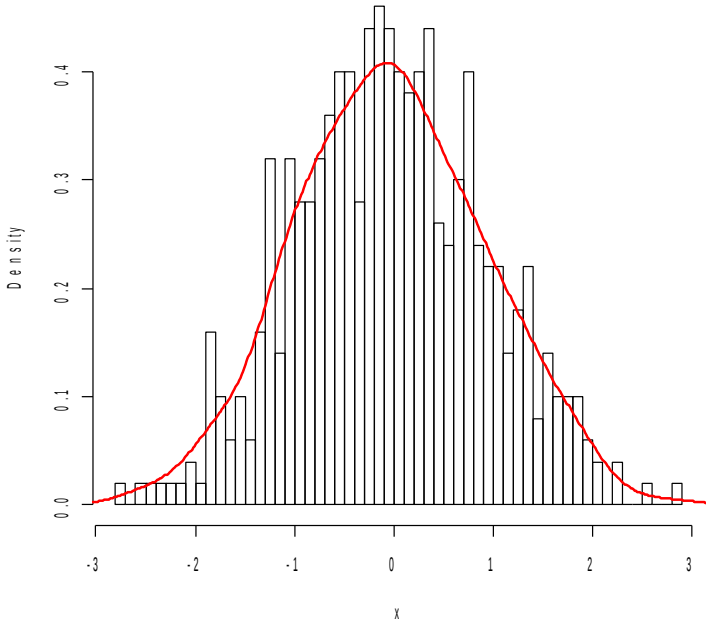
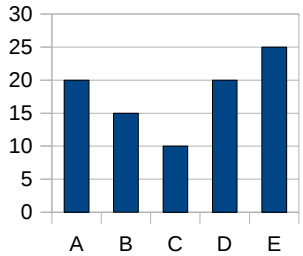
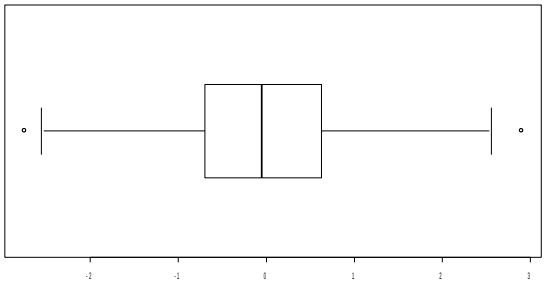
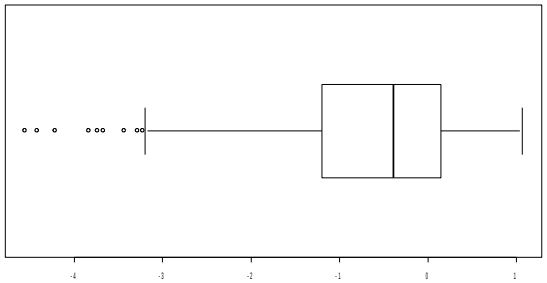
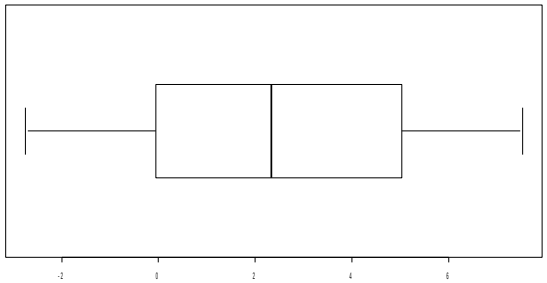
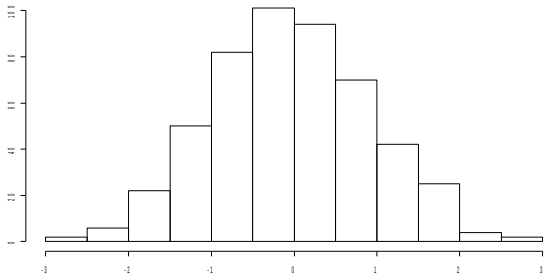
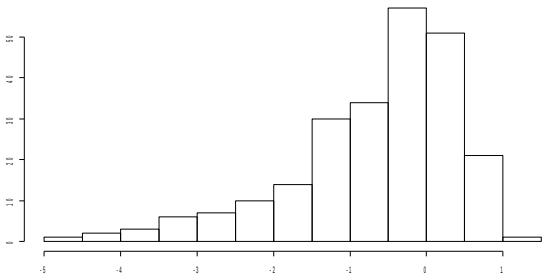
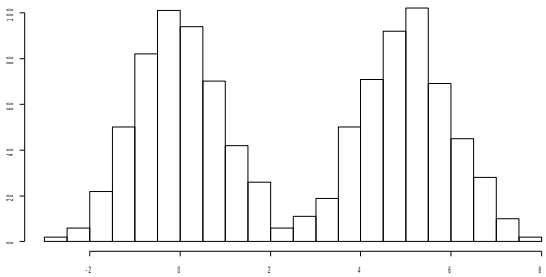


Diagramme en bâtons



Boxplot et histogramme

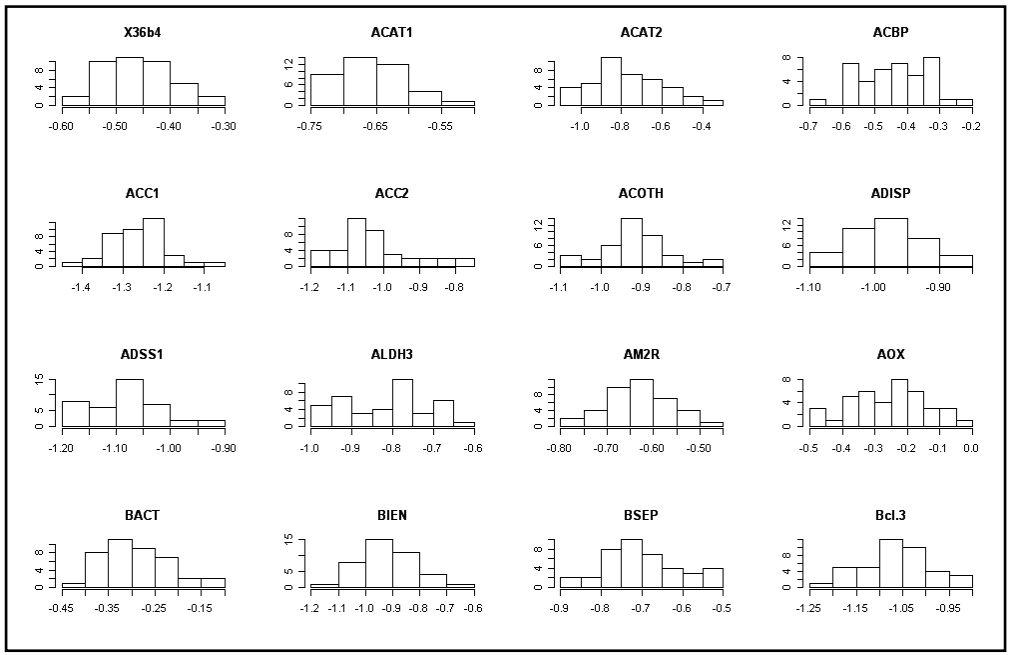







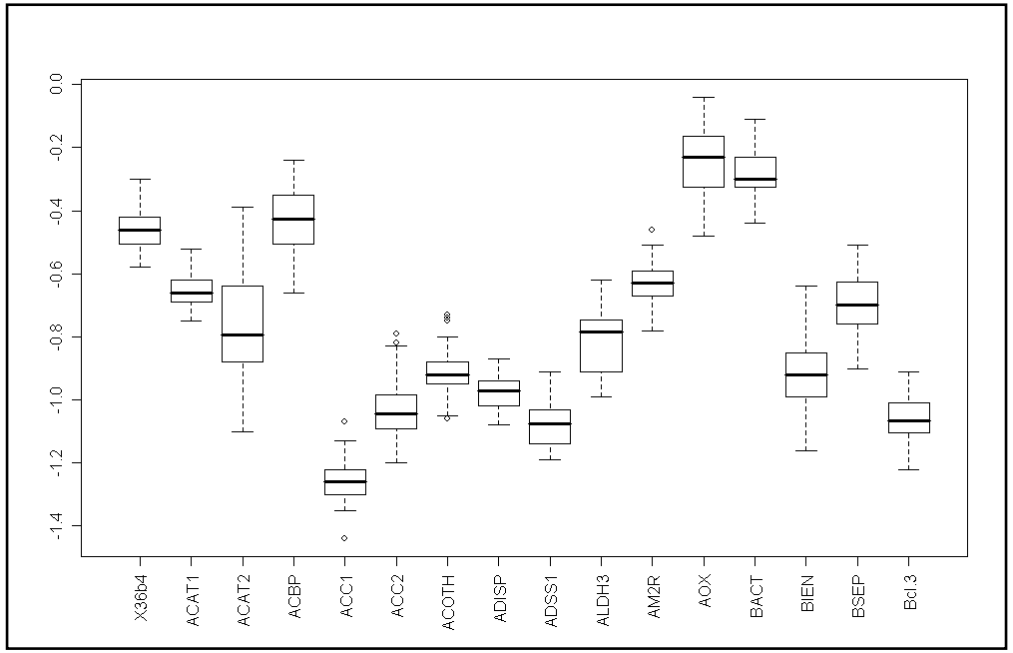
Individuellement, ces graphiques apportent la même information sur la position et la répartition des données. Attention cependant aux cas de multi-modalités que le boxplot ne peut pas capter.

⇒ ***avantage histogramme ?***

Sébastien Déjean - math.univ-toulouse.fr/~sdejean

Boxplot et histogramme



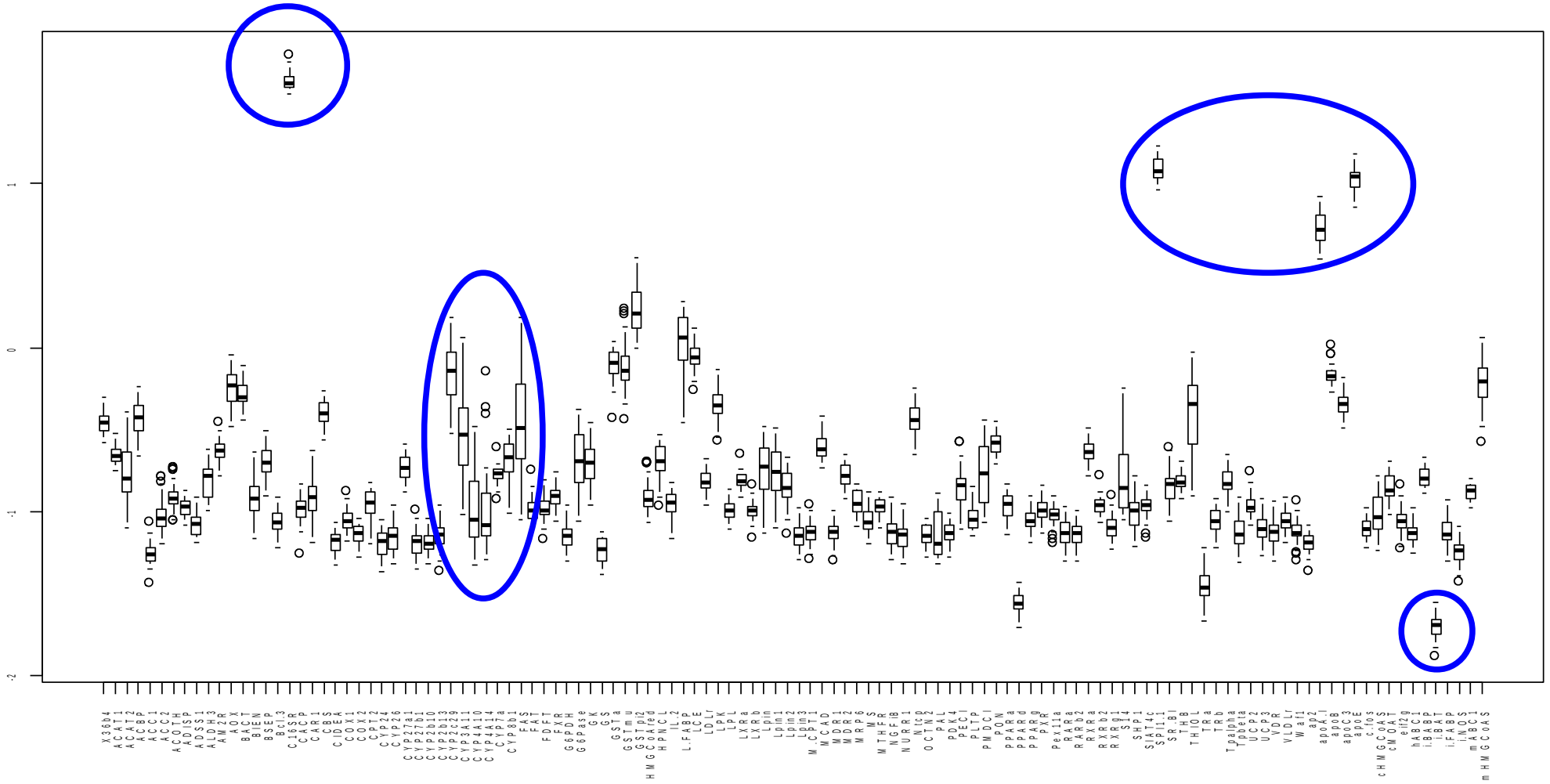


⇒ ***égalité ?***

B

Interpréter des *boxplot*

120 boxplots correspondant à l'expression de gènes mesurée sur **40** souris (**2** génotypes × **5** régimes alimentaires × **4** réplicats).

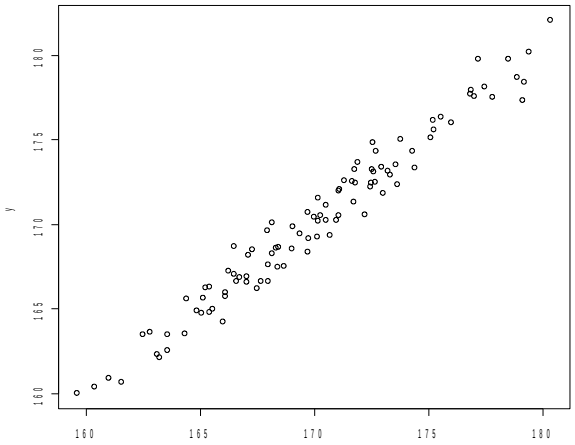


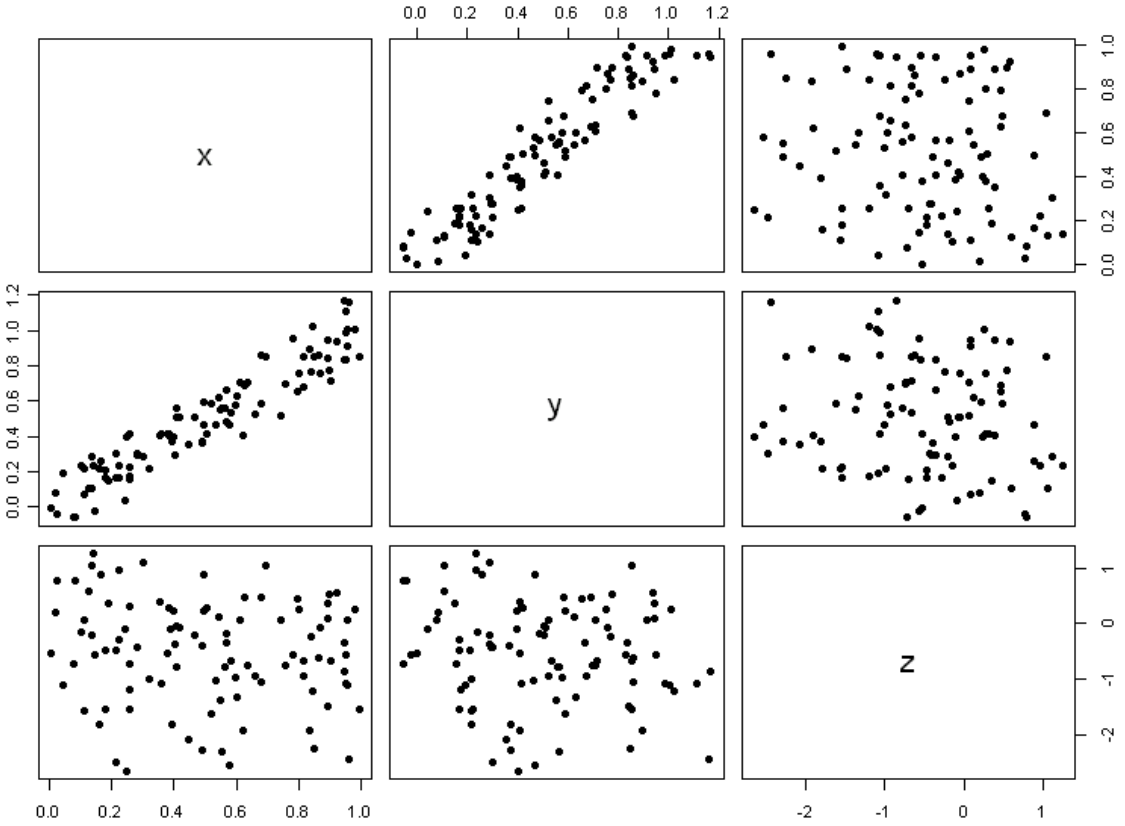
B

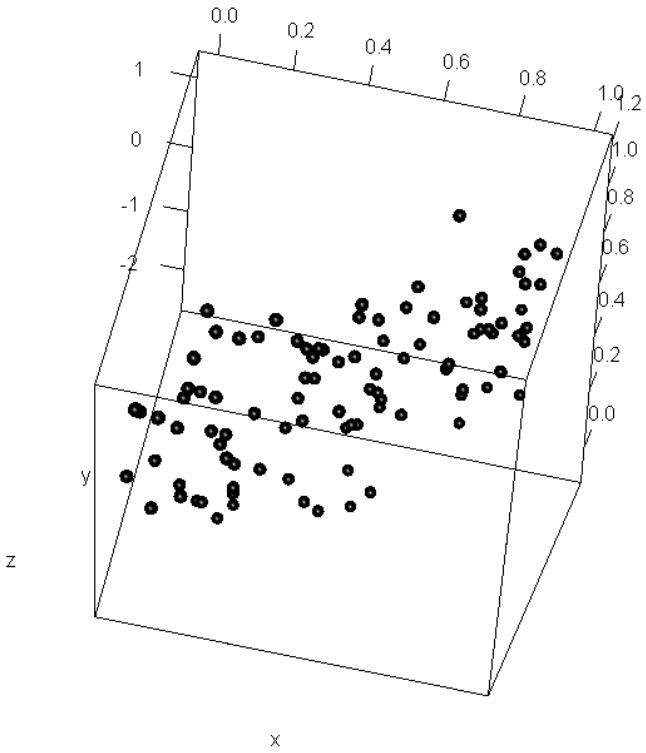
Représentations graphiques

2 variables ou plus

Nuage de points, diagramme de dispersion (*scatterplot*)

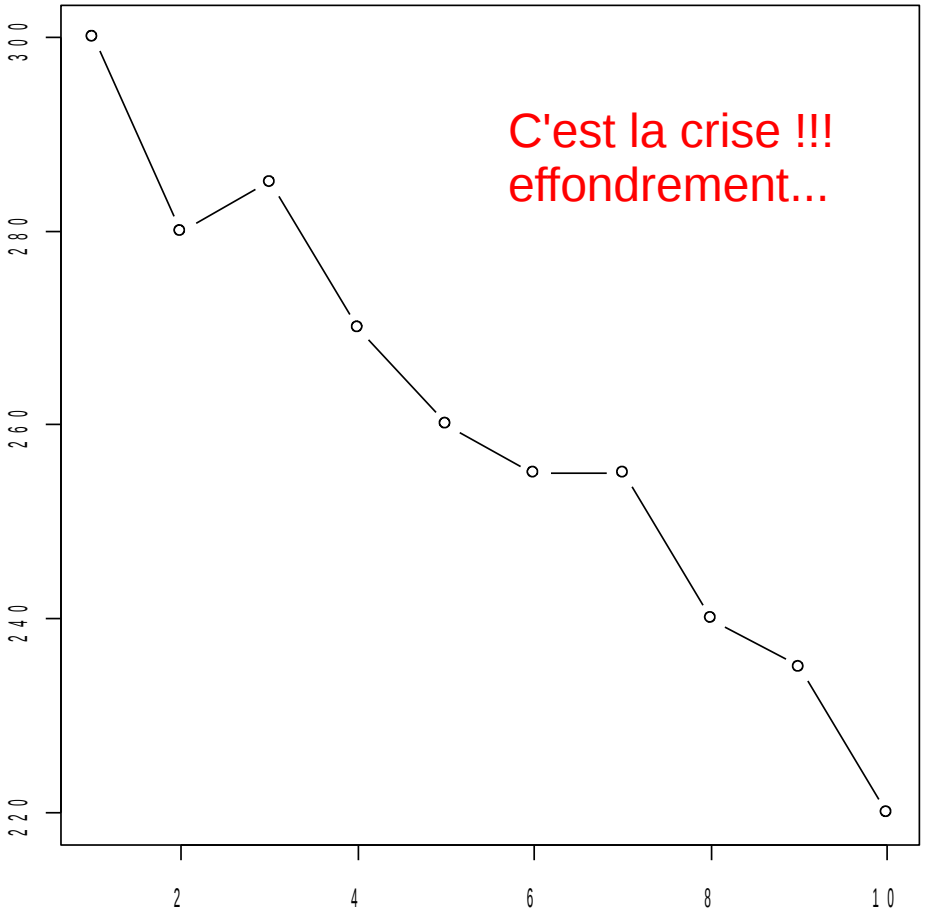


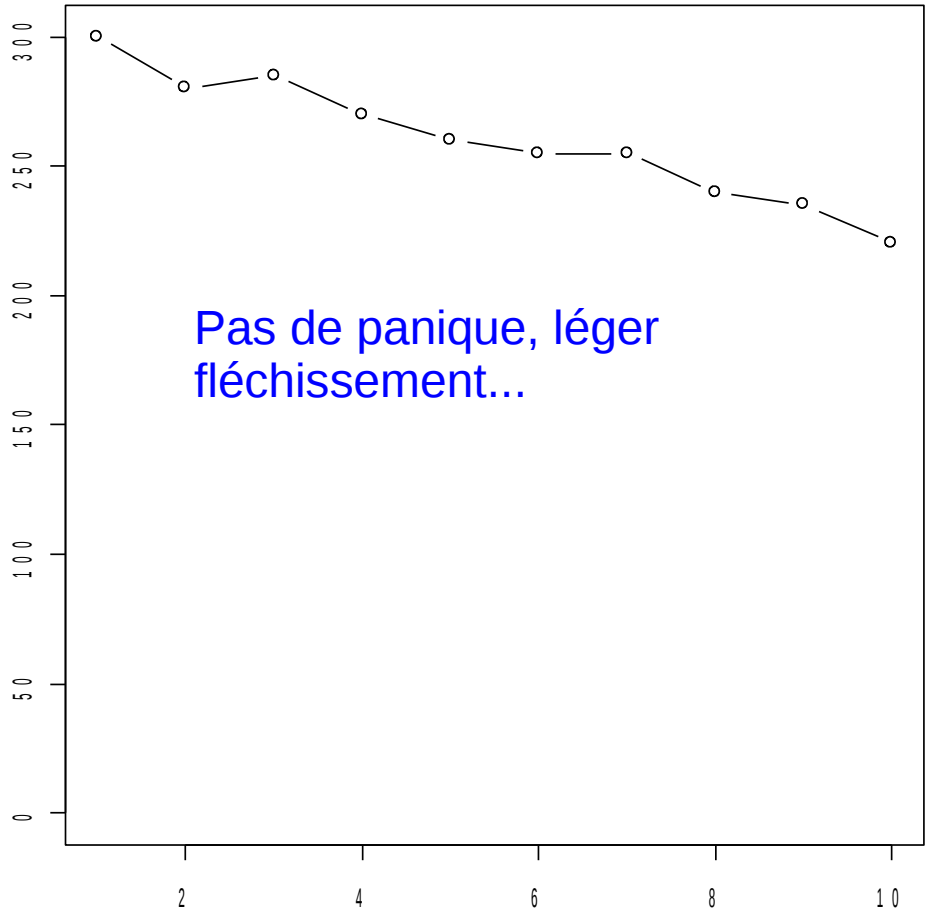




Échelles des axes

Temps	1	2	3	4	5	6	7	8	9	10
Chômage, CAC40, chiffre d'affaires...	300	280	285	270	260	255	255	240	235	220





Notions de probabilité

Probable : qui a une apparence de vérité ; dont la vérité a plus de raisons d'être confirmée que d'être infirmée. Qu'il est raisonnable de supposer, de conjecturer, de prévoir; qui a beaucoup de chances de se produire.

- « Quantifier » des événements dont la réalisation est incertaine

- « Formule » : nombre de cas favorables / nombre de cas possibles
- Exemple classique : jeu de « pile ou face »

  - Nombre de cas possibles : 2 (pile et face)
  - Nombre de cas favorables : 1 (le côté sur lequel on a misé)
  - Probabilité de gagner (ou de perdre) : 1/2

Espérance

Espérance : disposition de l'âme qui porte l'homme à considérer dans l'avenir un bien important qu'il désire et qu'il croit pouvoir se réaliser.

- Espérance mathématique : Valeur moyenne théorique d'une variable aléatoire

$$E(x) = x1.P(X=x1) + x2.P(X=x2) + ...$$

Exemples :

- Gain de 10€ si le dé indique une valeur paire 0€ sinon :

Espérance de gain : 10*3/6 + 0*3/6 = 5€

- Gain de 20€ si le dé tombe sur 6 et 10€ sur 5, 0 sinon

Espérance de gain : 20*1/6 + 10*1/6 + 0*4/6= 5€

Gagner peu souvent ⇔ gagner beaucoup rarement.

Variable aléatoire

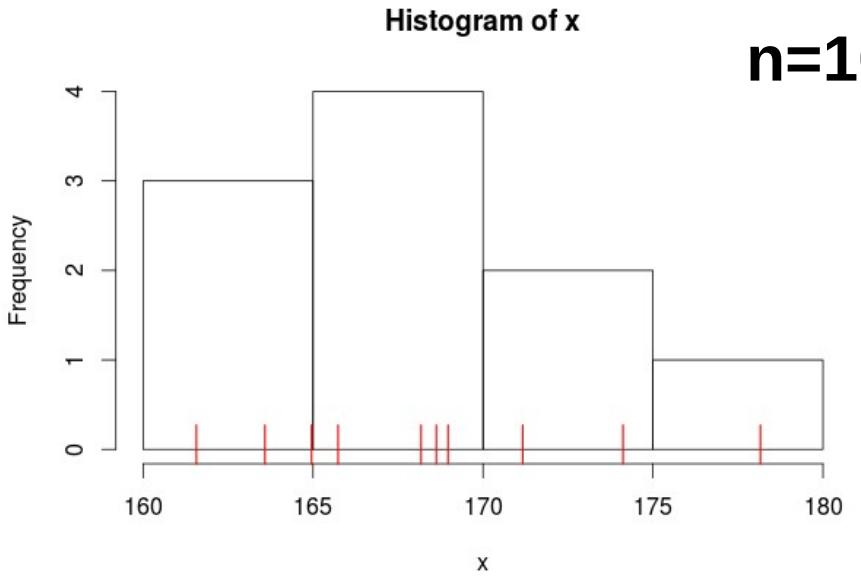
Une **variable aléatoire** est utilisée pour modéliser le résultat d'une expérience non déterministe qui génère un résultat aléatoire.

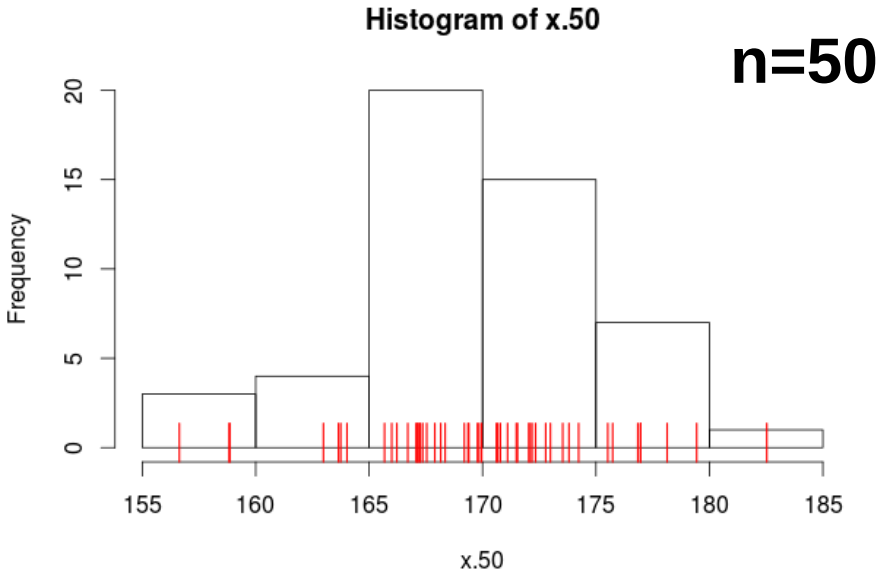
Exemples :

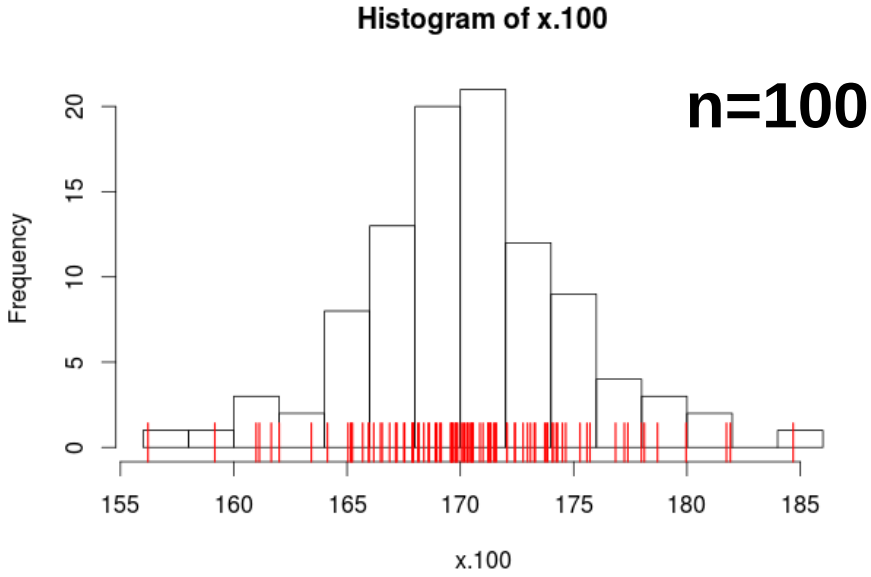
- lancement d'une pièce, d'un dé (v.a. **discrète**)
- taille d'un individu pris au hasard dans une population (v.a. **continue**)

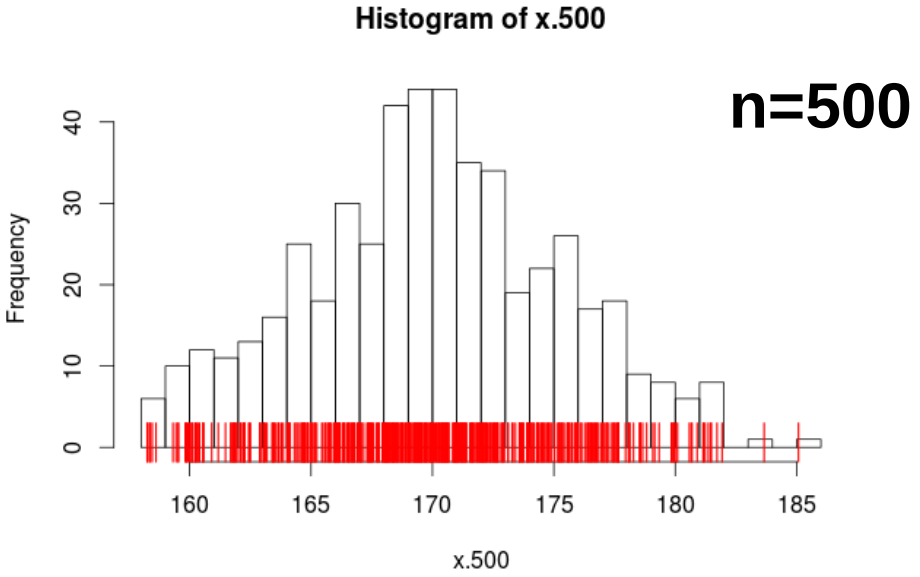
La répartition des valeurs prises par une v.a. conduit à la notion de **loi de probabilité**.

Exemple 1 : taille d'individus

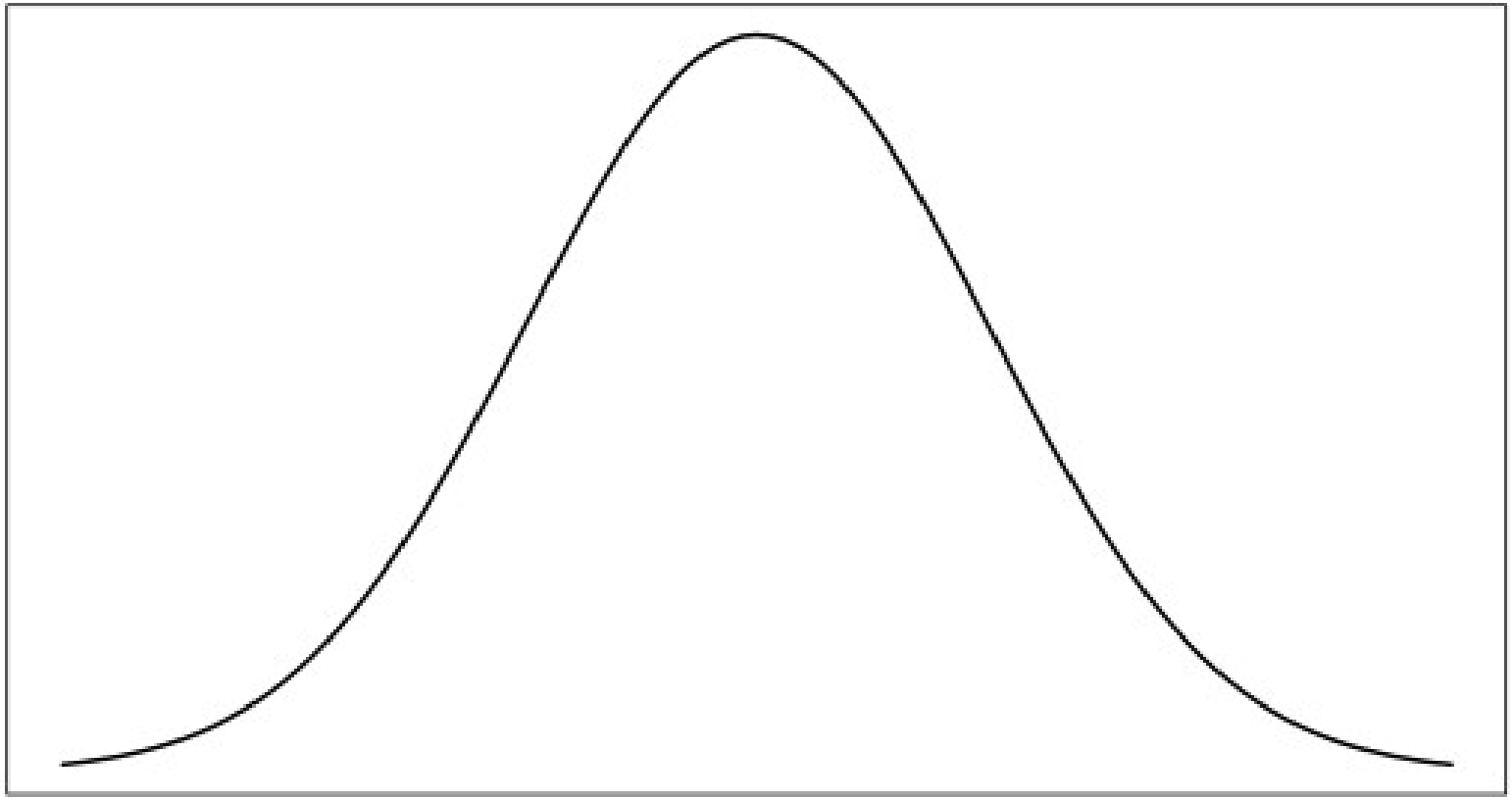




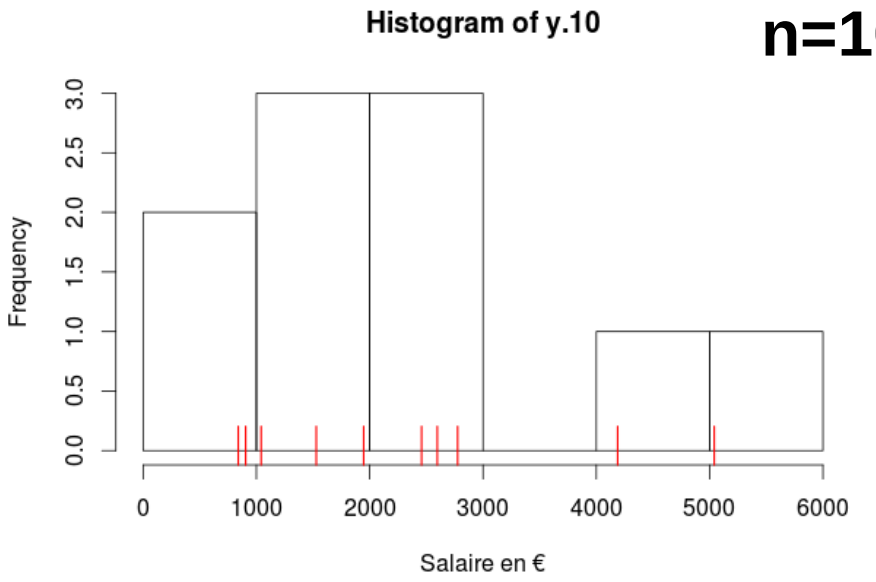


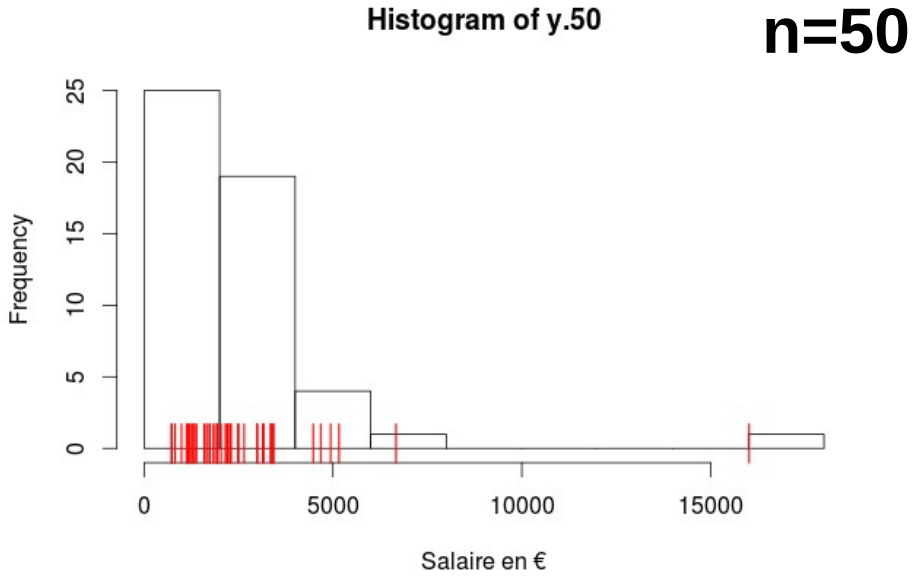


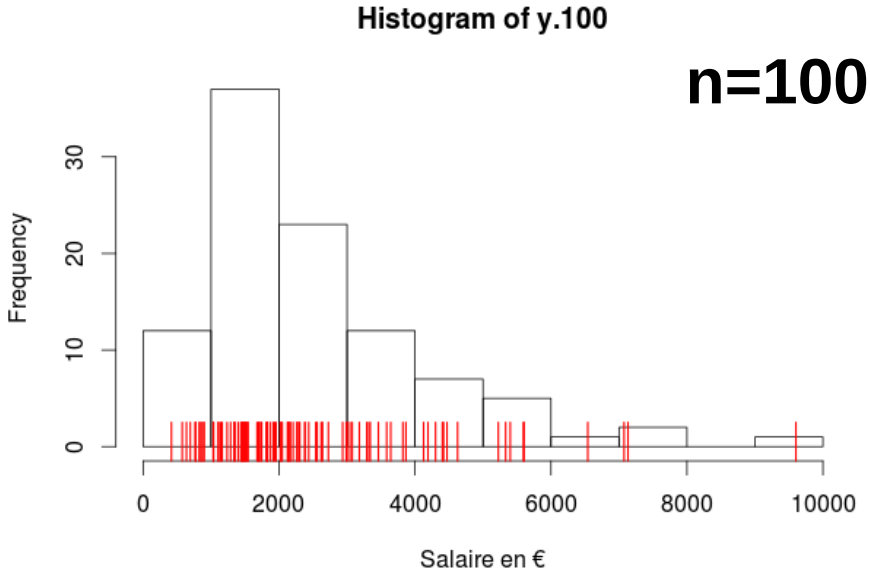
Loi normale, loi de Student

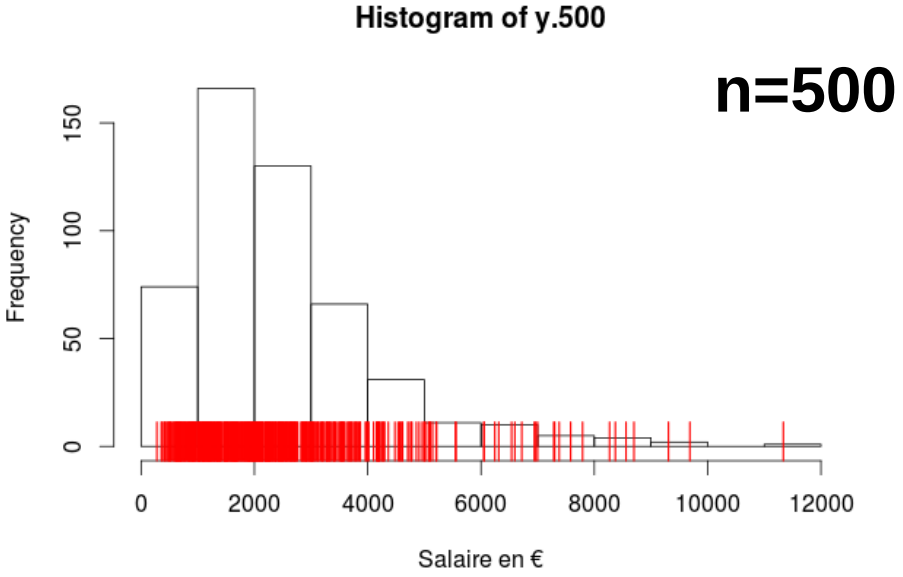


Exemple 2 : salaires

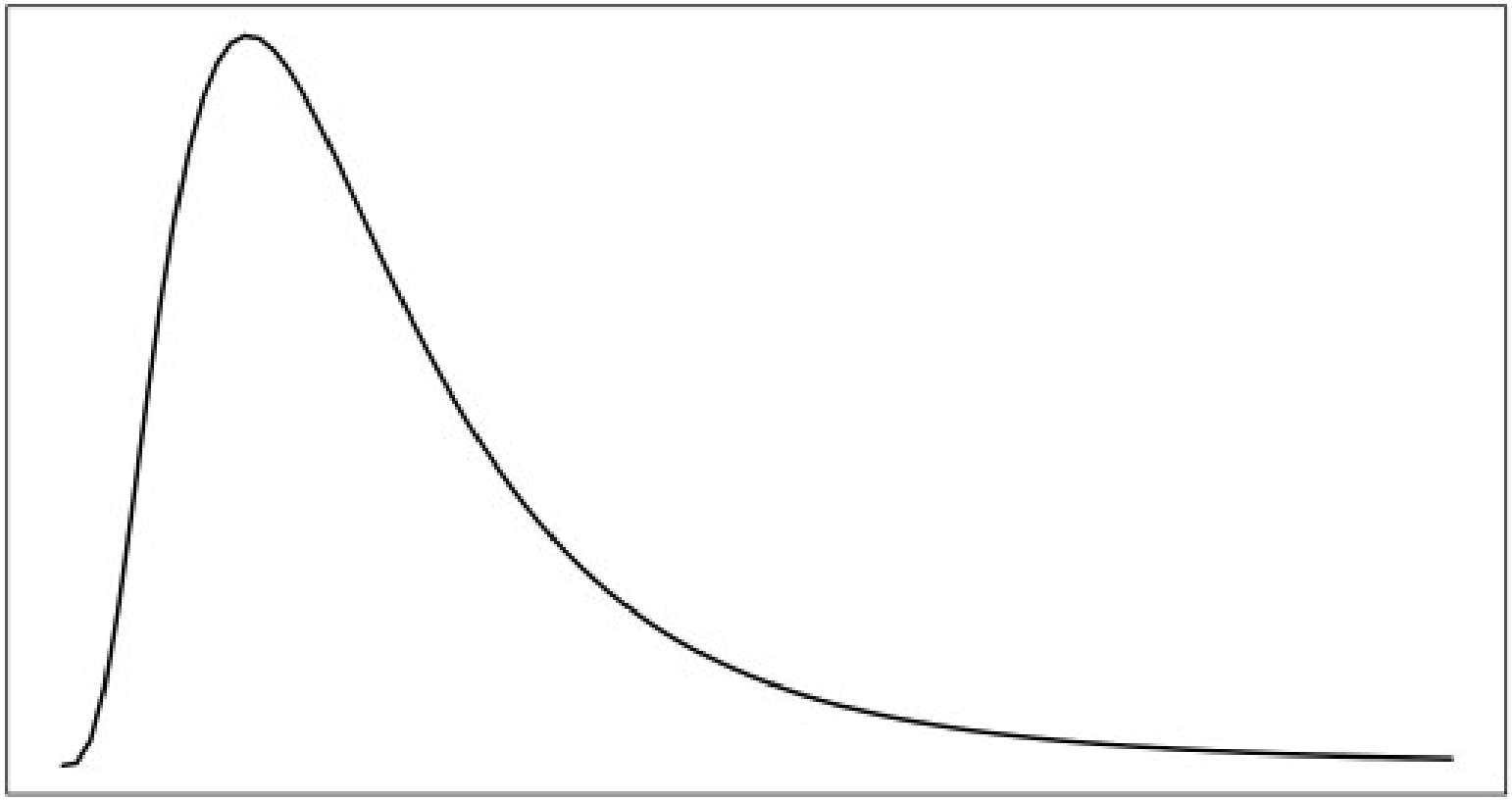






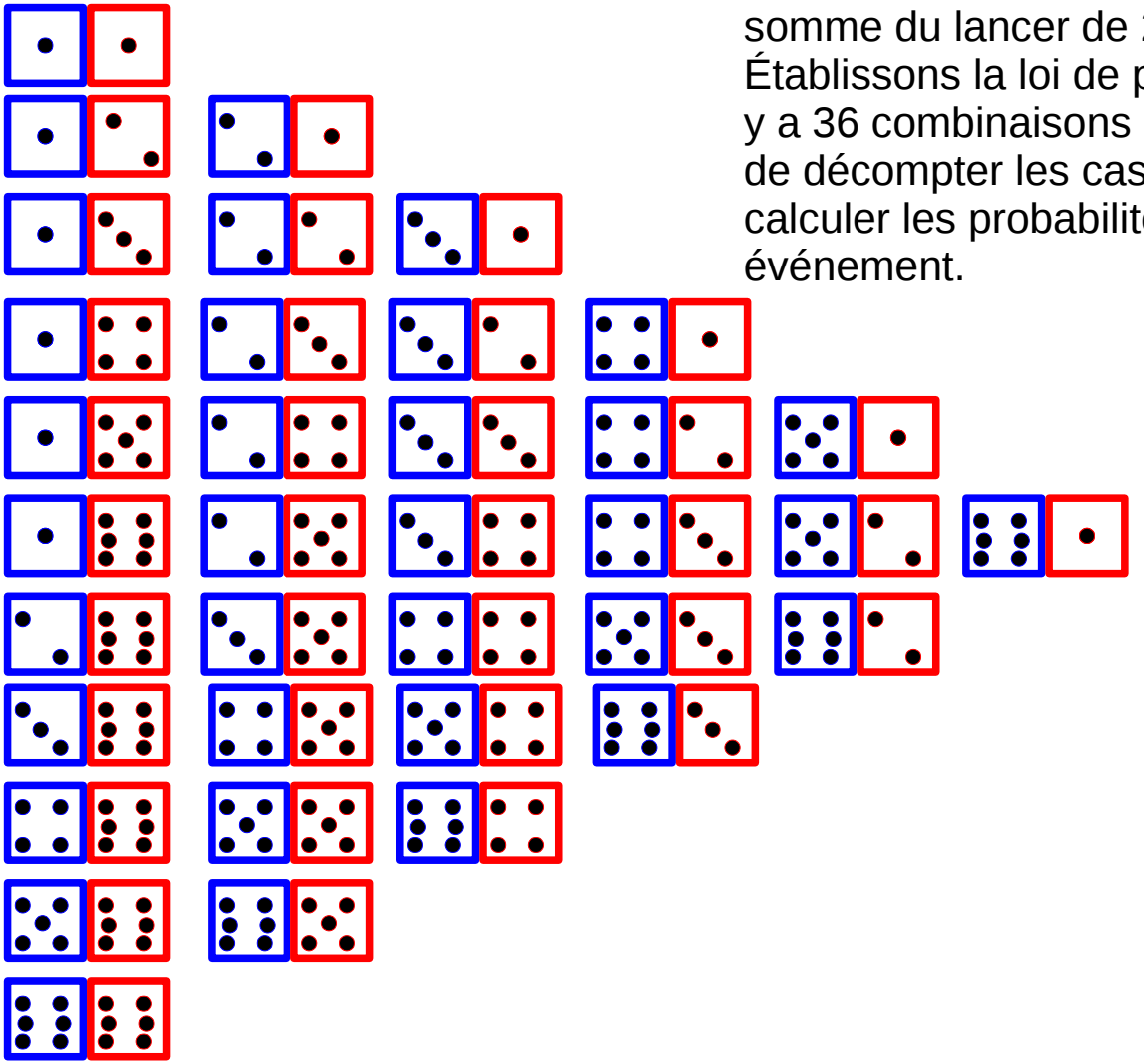


Loi de χ^2, loi de Fisher



Exemple 3 : lancer de 2 dés

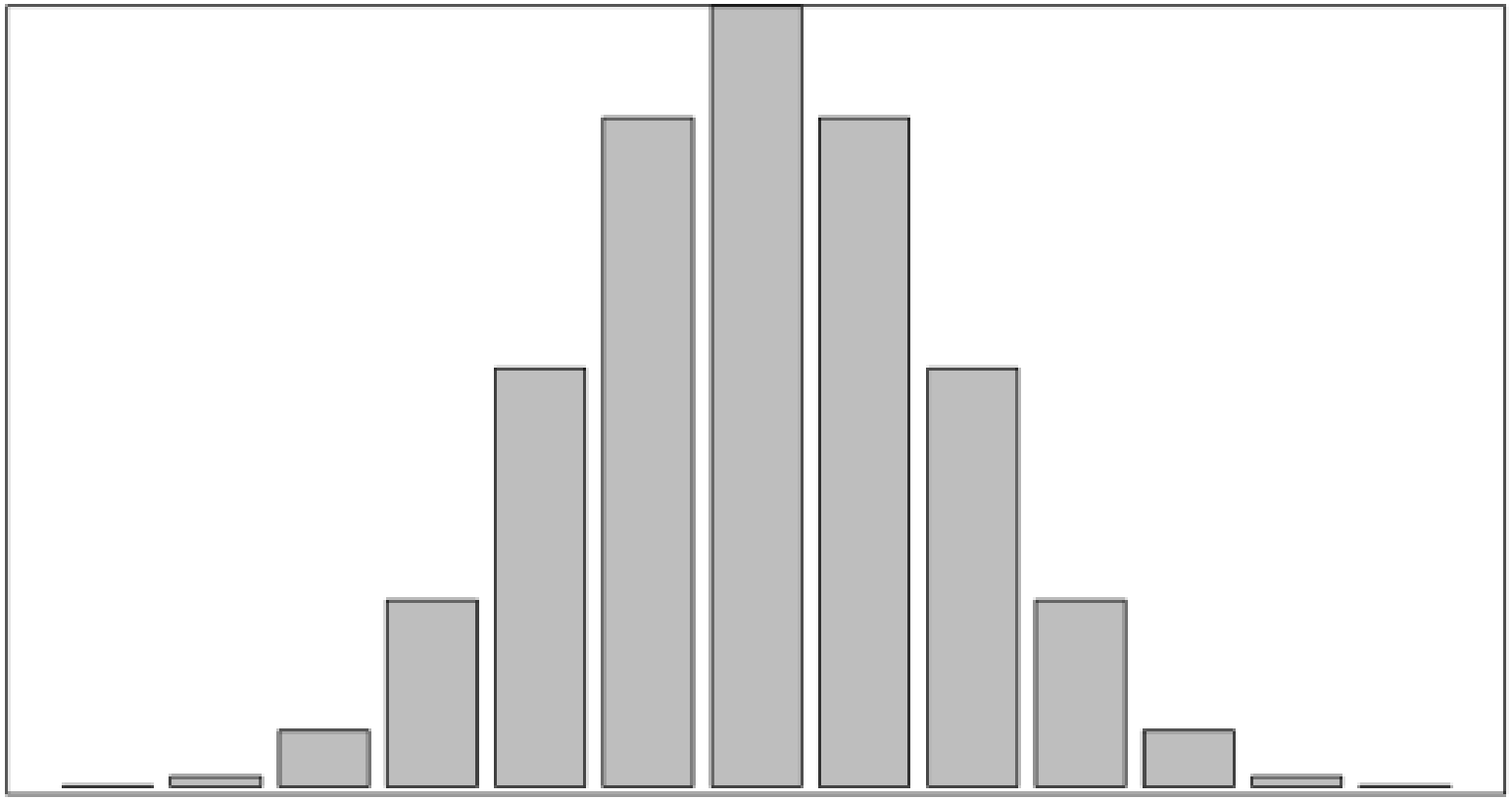
- P(X<0) = P(X=0) = P(X=1) = 0
- P(x=2) = 1/36
- P(x=3) = 2/36
- P(X=4) = 3/36
- P(X=5) = 4/36
- P(x=6) = 5/36
- P(X=7) = 6/36
- P(X=8) = 5/36
- P(X=9) = 4/36
- P(X=10) = 3/36
- P(X=11) = 2/36
- P(X=12) = 1/36
- P(X>12) = 0



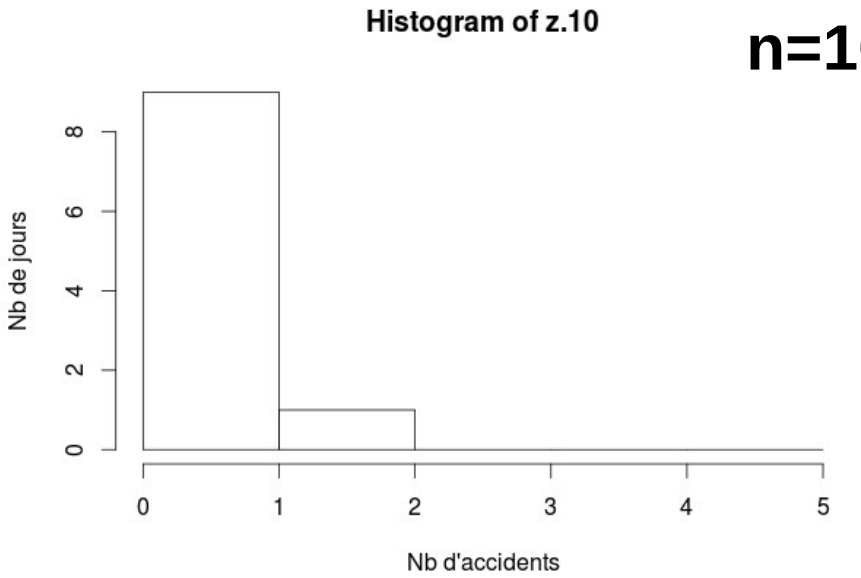
Soit X la variable aléatoire traduisant la somme du lancer de 2 dés à 6 faces. Établissons la loi de probabilité de X. Il y a 36 combinaisons possibles, il suffit de décompter les cas favorables pour calculer les probabilités de chaque événement.

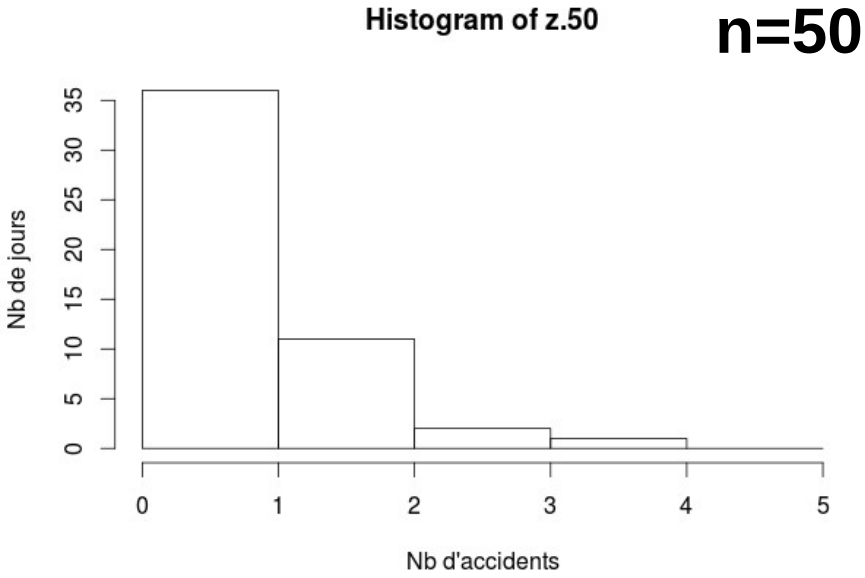
C

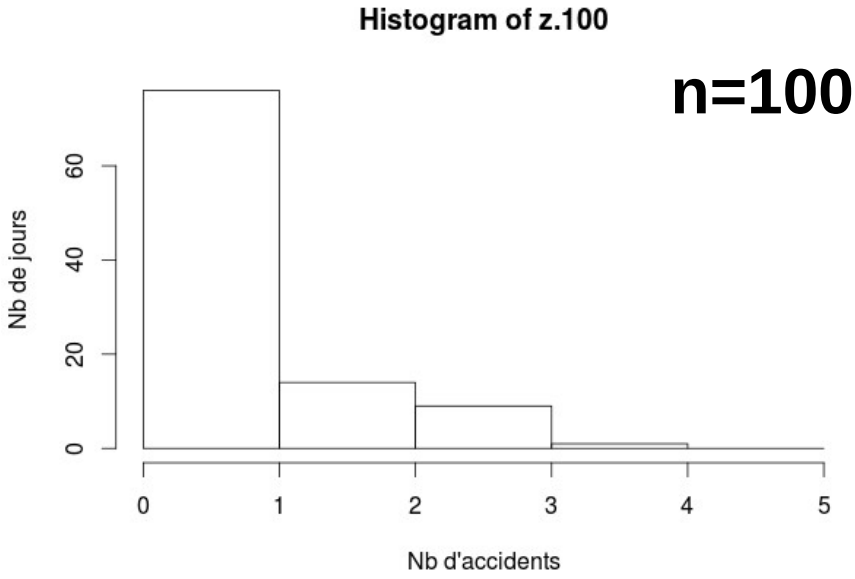
Loi binomiale

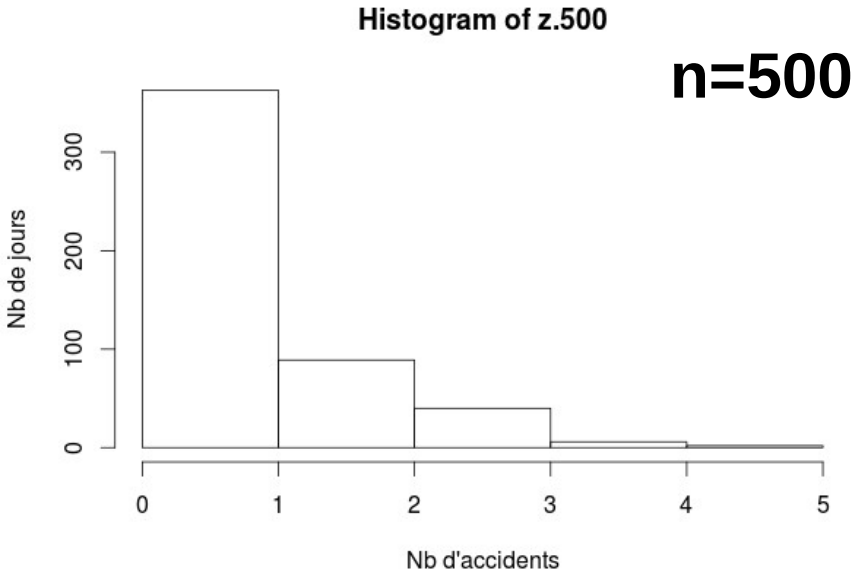


Exemple 4 : accidents

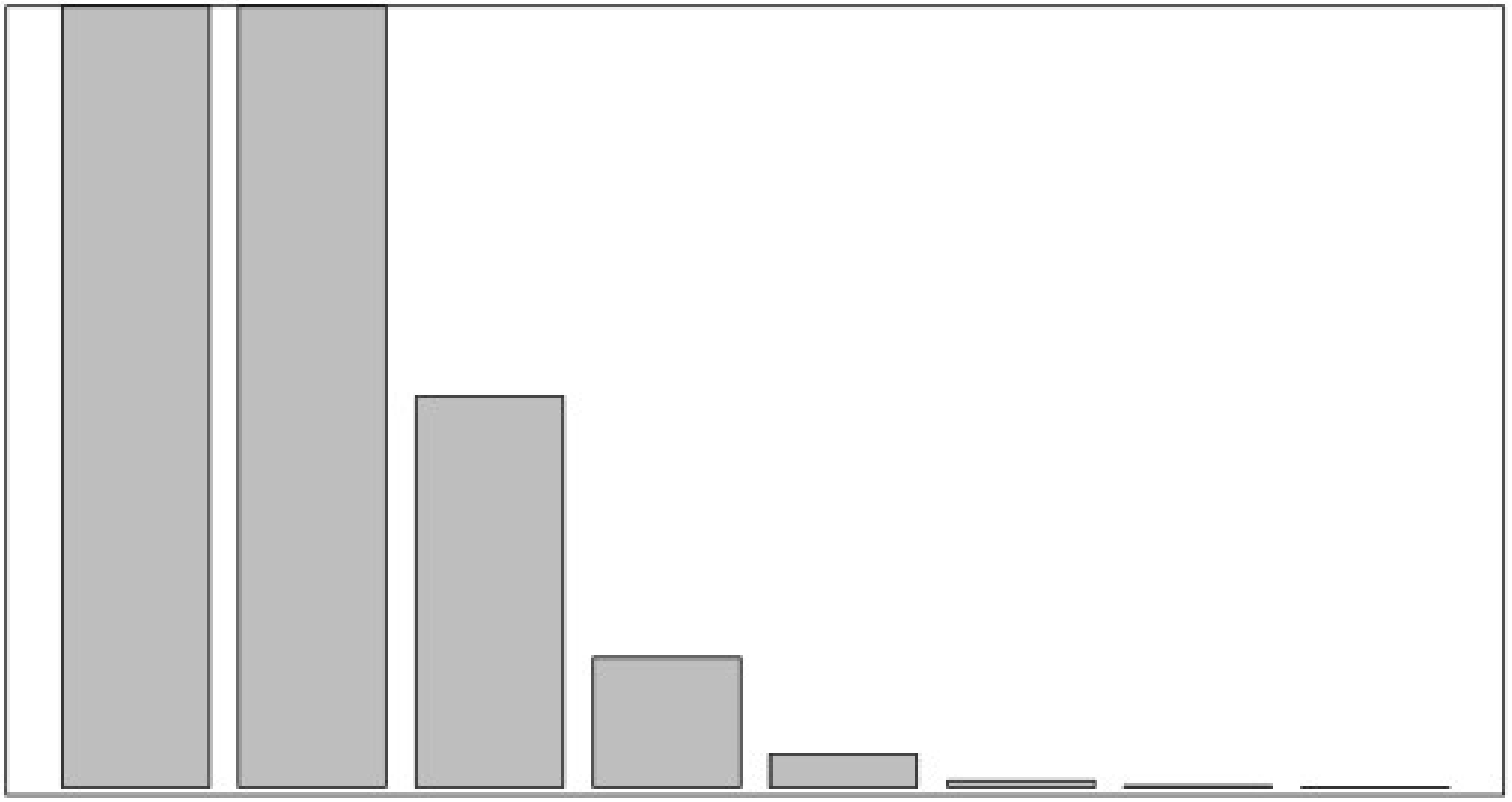








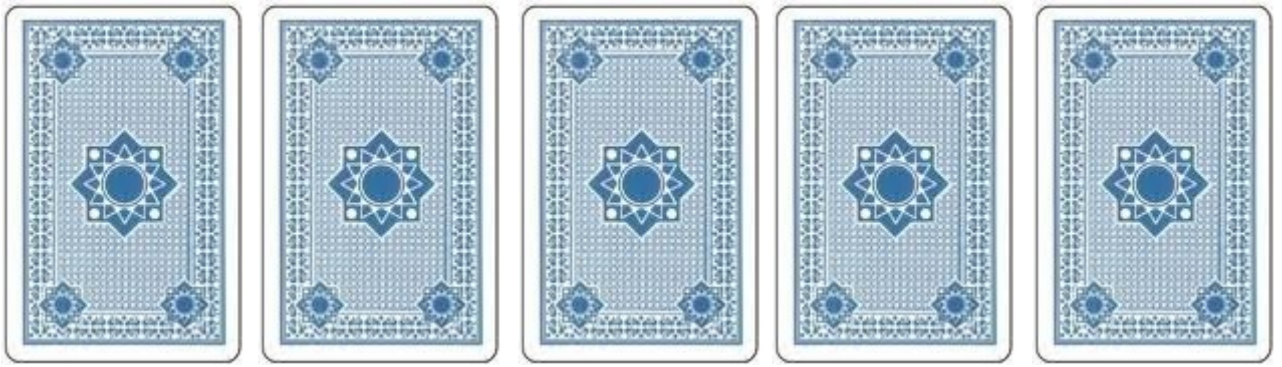
Loi de Poisson, loi binomiale négative



Exemple 5 : poker (simplifié)

Jeu de 52 cartes

Chaque joueur reçoit 5 cartes



Quelle est la probabilité
d'obtenir un carré d'as ?

Loi hypergéométrique

Probabilité d'avoir 0 As dans une main de 5 cartes alors qu'il y en 4 parmi 52 (48+4).

```
> dhyper(0, 4, 48, 5)
```

0.6588

Probabilité d'avoir 1 As dans une main de 5 cartes alors qu'il y en 4 parmi 52 (48+4).

```
> dhyper(1, 4, 48, 5)
```

0.2995

Probabilité d'avoir 2 As dans une main de 5 cartes alors qu'il y en 4 parmi 52 (48+4).

```
> dhyper(2, 4, 48, 5)
```

0.0399

Probabilité d'avoir 3 As dans une main de 5 cartes alors qu'il y en 4 parmi 52 (48+4).

```
> dhyper(3, 4, 48, 5)
```

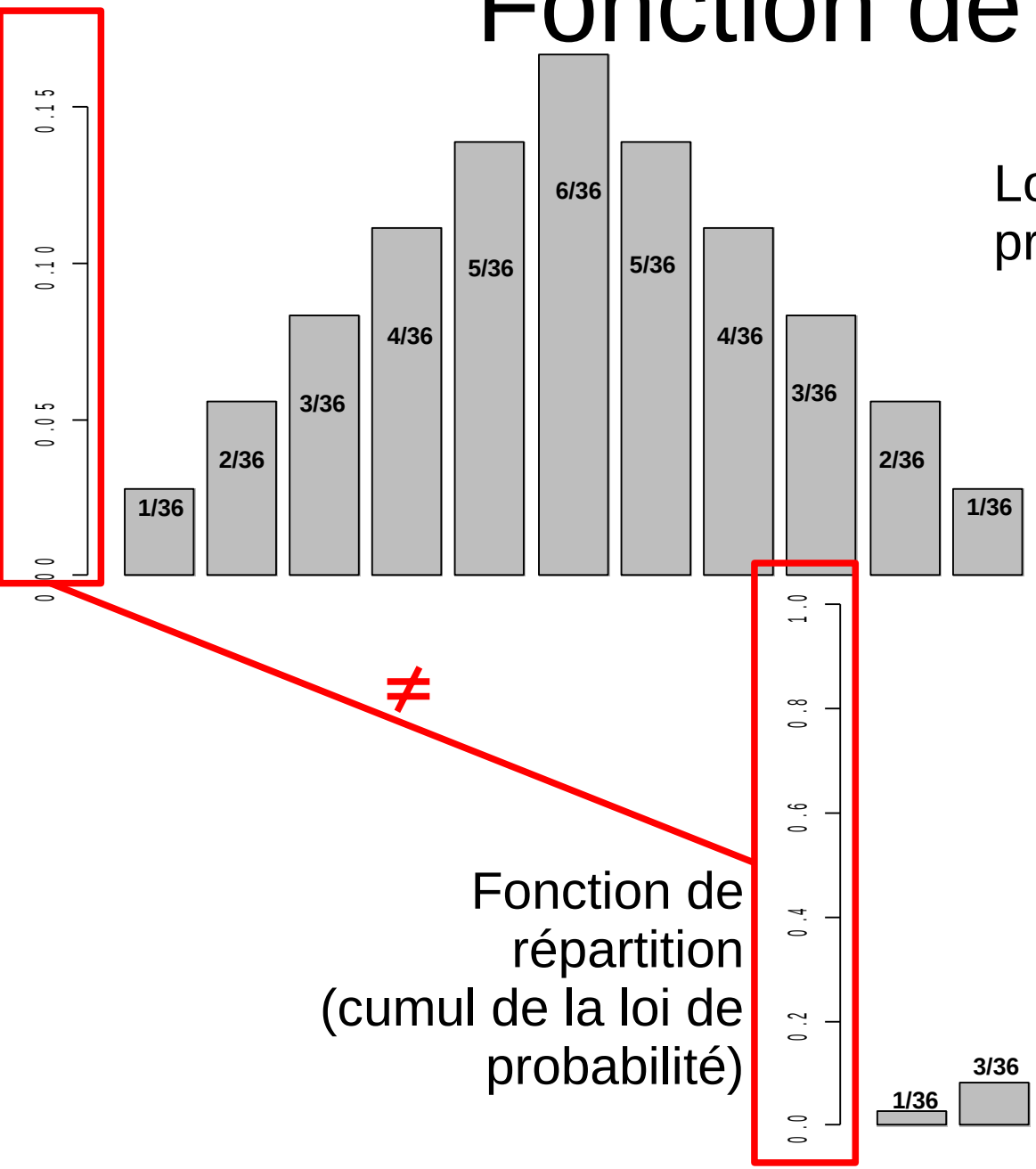
0.0017

Probabilité d'avoir 4 As dans une main de 5 cartes alors qu'il y en 4 parmi 52 (48+4).

```
> dhyper(4, 4, 48, 5)
```

0.00002 *2 chances sur 100 000*

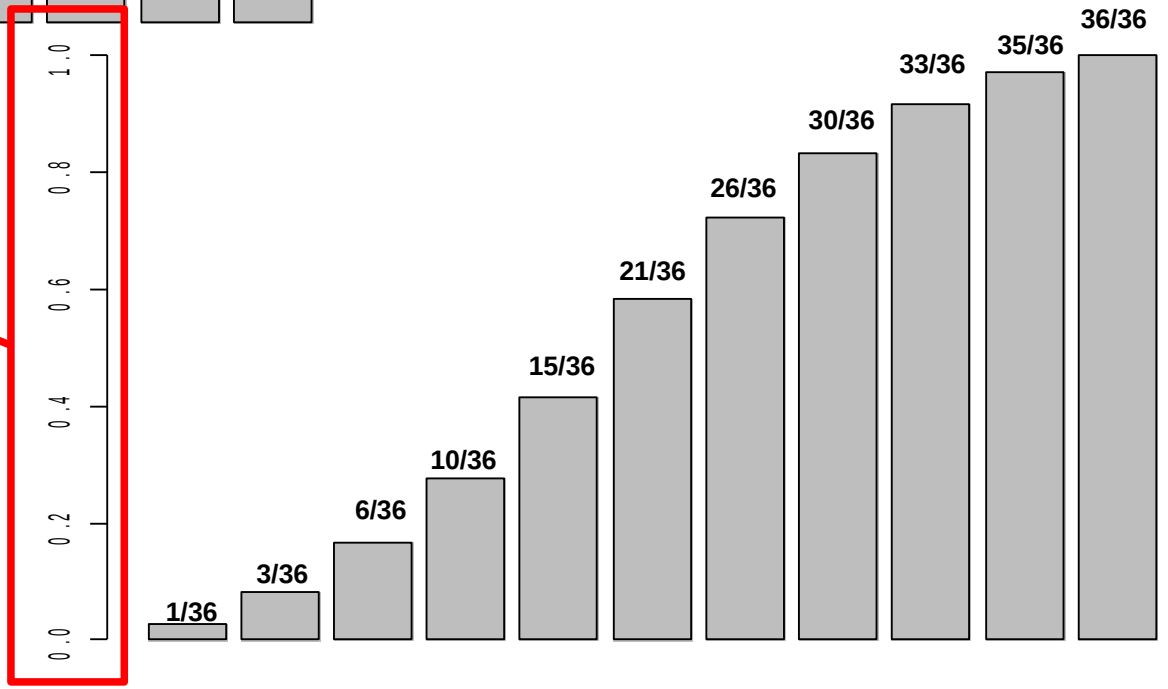
Distribution de probabilité – Fonction de répartition



Loi de probabilité

≠

Fonction de répartition (cumul de la loi de probabilité)



C

Distribution de probabilité – Fonction de répartition

Loi de probabilité

- dérivée de la fonction de répartition
- à valeur positive ou nulle

Fonction de répartition

- croissante sur]-∞;+∞[
- tend vers 0 en -∞ et 1 en +∞
- continue à droite en tout point

Cas discret

$$P_X(k)=P(X=k)$$

$$\sum_{k\in Z} P(X=k)=1$$

$$F_X(n)=\sum_{k=-\infty}^{n} P(X=k)$$

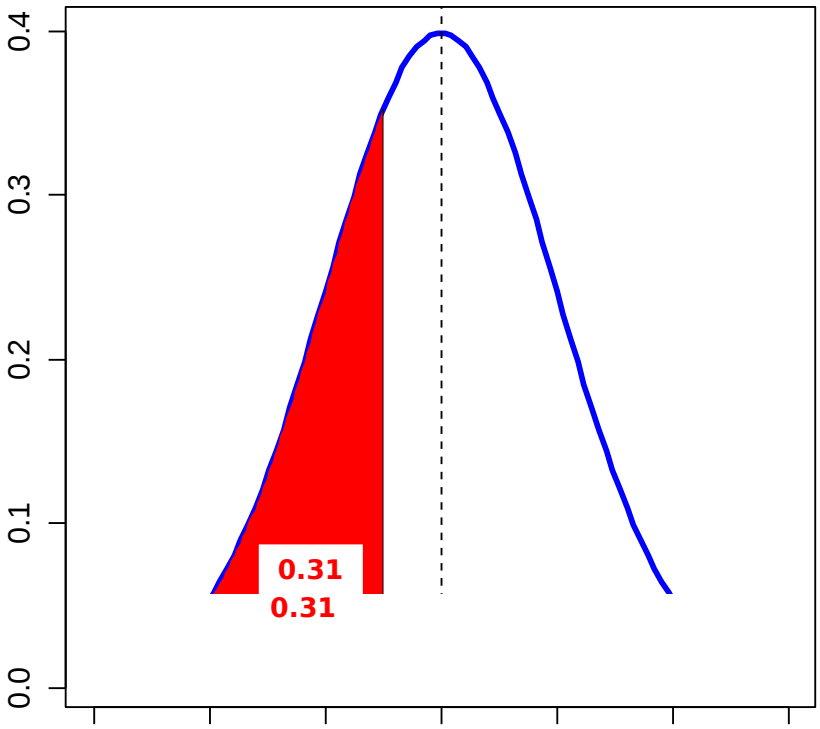
Cas continu (densité f)

$$P(a<X<b)=\int_a^b f(x)dx$$

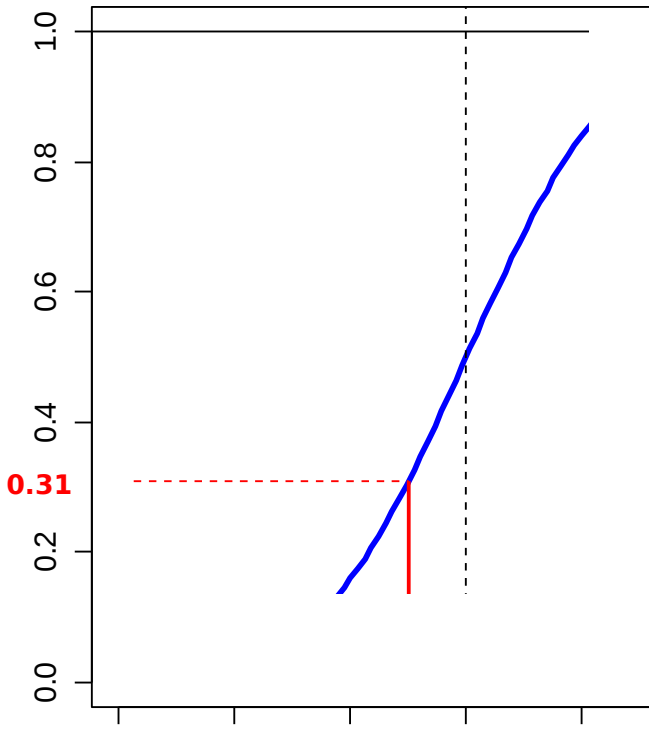
$$\int_{-\infty}^{+\infty} f(x)dx=1$$

$$F_X(x)=\int_{-\infty}^{x} f(t)dt$$

Distribution de probabilité – Fonction de répartition



Loi de probabilité



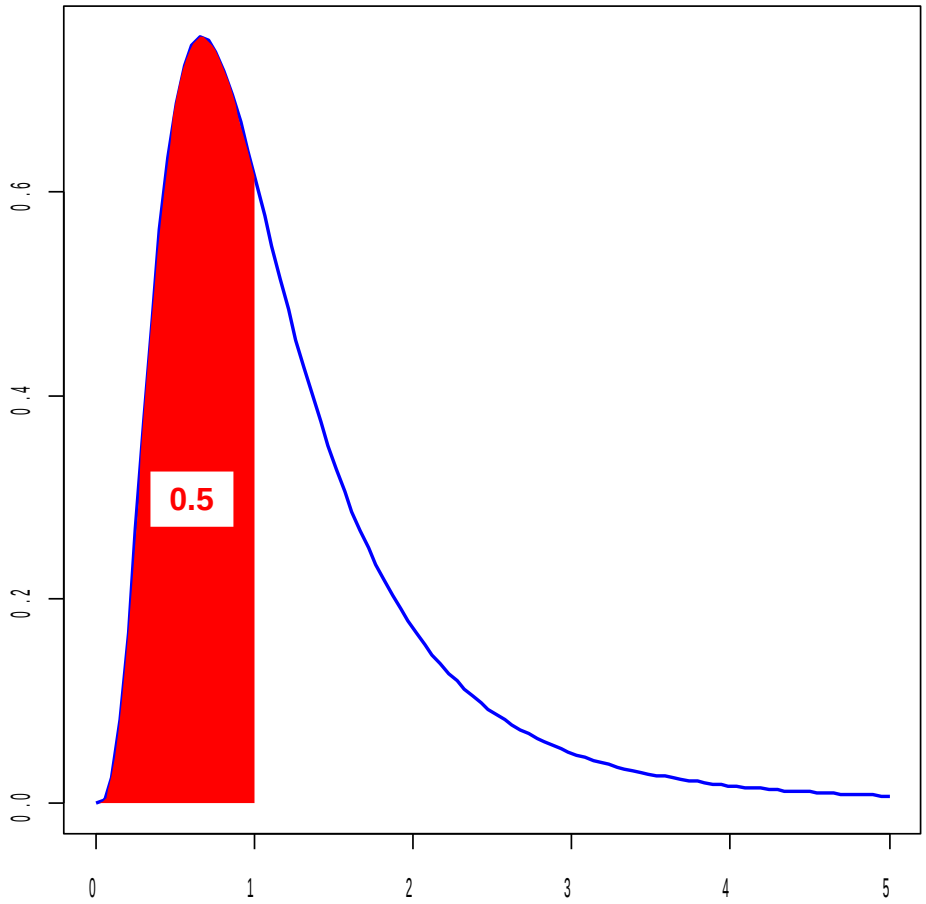
Fonction de répartition

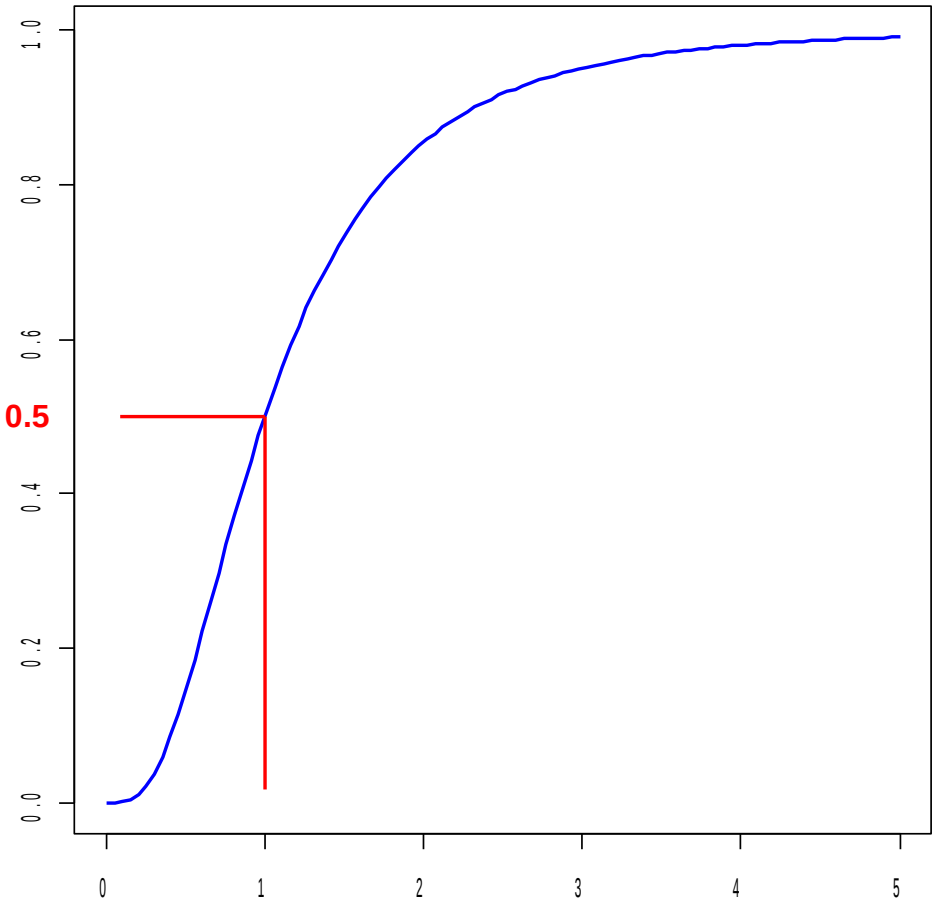
Aire sous la courbe = probabilité

C

Distribution de probabilité – Fonction de répartition

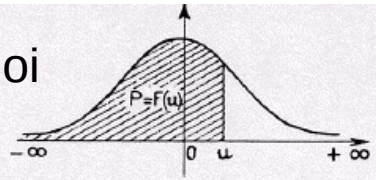
Exemple avec une loi asymétrique : F(10,10)





Histoire : les tables statistiques

Fonction de répartition de la loi normale N(0,1)

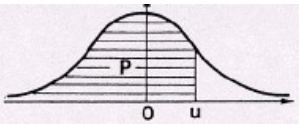


u	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7290	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9779	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986

Table pour les grandes valeurs de u

u	3,0	3,1	3,2	3,3	3,4	3,5	3,6	3,8	4,0	4,5
F(u)	0,99865	0,99904	0,99931	0,99952	0,99966	0,99976	0,999841	0,999928	0,999968	0,999997

Fractiles de la loi normale N(0,1)



P	0,000	0,001	0,002	0,003	0,004	0,005	0,006	0,007	0,008	0,009	0,010	
0,00	∞	3,0902	2,8782	2,7478	2,6521	2,5758	2,5121	2,4573	2,4089	2,3656	2,3263	0,99
0,01	2,3263	2,2904	2,2571	2,2262	2,1973	2,1701	2,1444	2,1201	2,0969	2,0749	2,0537	0,98
0,02	2,0537	2,0335	2,0141	1,9954	1,9774	1,9600	1,9431	1,9268	1,9110	1,8957	1,8808	0,97
0,03	1,8808	1,8663	1,8522	1,8384	1,8250	1,8119	1,7991	1,7866	1,7744	1,7624	1,7507	0,96
0,04	1,7507	1,7392	1,7279	1,7169	1,7060	1,6954	1,6849	1,6747	1,6646	1,6546	1,6449	0,95
0,05	1,6449	1,6352	1,6258	1,6164	1,6072	1,5982	1,5893	1,5805	1,5718	1,5632	1,5548	0,94
0,06	1,5548	1,5464	1,5382	1,5301	1,5220	1,5141	1,5063	1,4985	1,4909	1,4833	1,4758	0,93
0,07	1,4758	1,4684	1,4611	1,4538	1,4466	1,4395	1,4325	1,4255	1,4187	1,4118	1,4051	0,92
0,08	1,4051	1,3984	1,3917	1,3852	1,3787	1,3722	1,3658	1,3595	1,3532	1,3469	1,3408	0,91
0,09	1,3408	1,3346	1,3285	1,3225	1,3165	1,3106	1,3047	1,2988	1,2930	1,2873	1,2816	0,90
0,10	1,2816	1,2759	1,2702	1,2646	1,2591	1,2536	1,2481	1,2426	1,2372	1,2319	1,2265	0,89
0,11	1,2265	1,2212	1,2160	1,2107	1,2055	1,2004	1,1952	1,1901	1,1850	1,1800	1,1750	0,88
0,12	1,1750	1,1700	1,1650	1,1601	1,1552	1,1503	1,1455	1,1407	1,1359	1,1311	1,1264	0,87
0,13	1,1264	1,1217	1,1170	1,1123	1,1077	1,1031	1,0985	1,0939	1,0893	1,0848	1,0803	0,86
0,14	1,0803	1,0758	1,0714	1,0669	1,0625	1,0581	1,0537	1,0494	1,0450	1,0407	1,0364	0,85
0,15	1,0364	1,0322	1,0279	1,0237	1,0194	1,0152	1,0110	1,0069	1,0027	0,9986	0,9945	0,84
0,16	0,9945	0,9904	0,9863	0,9822	0,9782	0,9741	0,9701	0,9661	0,9621	0,9581	0,9542	0,83
0,17	0,9542	0,9502	0,9463	0,9424	0,9385	0,9346	0,9307	0,9269	0,9230	0,9192	0,9154	0,82
0,18	0,9154	0,9116	0,9078	0,9040	0,9002	0,8965	0,8927	0,8890	0,8853	0,8816	0,8779	0,81
0,19	0,8779	0,8742	0,8705	0,8669	0,8633	0,8596	0,8560	0,8524	0,8488	0,8452	0,8416	0,80
0,20	0,8416	0,8381	0,8345	0,8310	0,8274	0,8239	0,8204	0,8169	0,8134	0,8099	0,8064	0,79
0,21	0,8064	0,8030	0,7995	0,7961	0,7928	0,7892	0,7858	0,7824	0,7790	0,7756	0,7722	0,78
0,22	0,7722	0,7688	0,7655	0,7621	0,7588	0,7554	0,7521	0,7488	0,7454	0,7421	0,7388	0,77
0,23	0,7388	0,7356	0,7323	0,7290	0,7257	0,7225	0,7192	0,7160	0,7128	0,7095	0,7063	0,76
0,24	0,7063	0,7031	0,6999	0,6967	0,6935	0,6903	0,6871	0,6840	0,6808	0,6776	0,6745	0,75
0,25	0,6745	0,6713	0,6682	0,6651	0,6620	0,6588	0,6557	0,6526	0,6495	0,6464	0,6433	0,74
0,26	0,6433	0,6403	0,6372	0,6341	0,6311	0,6280	0,6250	0,6219	0,6189	0,6158	0,6128	0,73
0,27	0,6128	0,6098	0,6068	0,6038	0,6008	0,5978	0,5948	0,5918	0,5888	0,5858	0,5828	0,72
0,28	0,5828	0,5799	0,5769	0,5740	0,5710	0,5681	0,5651	0,5622	0,5592	0,5563	0,5534	0,71
0,29	0,5534	0,5505	0,5476	0,5446	0,5417	0,5388	0,5359	0,5330	0,5302	0,5273	0,5244	0,70
0,30	0,5244	0,5215	0,5187	0,5158	0,5129	0,5101	0,5072	0,5044	0,5015	0,4987	0,4959	0,69
0,31	0,4959	0,4930	0,4902	0,4874	0,4845	0,4817	0,4789	0,4761	0,4733	0,4705	0,4677	0,68
0,32	0,4677	0,4649	0,4621	0,4593	0,4565	0,4538	0,4510	0,4482	0,4454	0,4427	0,4399	0,67
0,33	0,4399	0,4372	0,4344	0,4316	0,4289	0,4261	0,4234	0,4207	0,4179	0,4152	0,4125	0,66
0,34	0,4125	0,4097	0,4070	0,4043	0,4016	0,3989	0,3961	0,3934	0,3907	0,3880	0,3853	0,65
0,35	0,3853	0,3826	0,3799	0,3772	0,3745	0,3719	0,3692	0,3665	0,3638	0,3611	0,3585	0,64
0,36	0,3585	0,3558	0,3531	0,3505	0,3478	0,3451	0,3425	0,3398	0,3372	0,3345	0,3319	0,63
0,37	0,3319	0,3292	0,3266	0,3239	0,3213	0,3186	0,3160	0,3134	0,3107	0,3081	0,3055	0,62
0,38	0,3055	0,3029	0,3002	0,2976	0,2950	0,2924	0,2898	0,2871	0,2845	0,2819	0,2793	0,61
0,39	0,2793	0,2767	0,2741	0,2715	0,2689	0,2663	0,2637	0,2611	0,2585	0,2559	0,2533	0,60
0,40	0,2533	0,2508	0,2482	0,2456	0,2430	0,2404	0,2378	0,2353	0,2327	0,2301	0,2275	0,59
0,41	0,2275	0,2250	0,2224	0,2198	0,2173	0,2147	0,2121	0,2096	0,2070	0,2045	0,2019	0,58
0,42	0,2019	0,1993	0,1968	0,1942	0,1917	0,1891	0,1866	0,1840	0,1815	0,1789	0,1764	0,57
0,43	0,1764	0,1738	0,1713	0,1687	0,1662	0,1637	0,1611	0,1586	0,1560	0,1535	0,1510	0,56
0,44	0,1510	0,1484	0,1459	0,1434	0,1408	0,1383	0,1358	0,1332	0,1307	0,1282	0,1257	0,55
0,45	0,1257	0,1231	0,1206	0,1181	0,1156	0,1130	0,1105	0,1080	0,1055	0,1030	0,1004	0,54
0,46	0,1004	0,0979	0,0954	0,0929	0,0904	0,0878	0,0853	0,0828	0,0803	0,0778	0,0753	0,53
0,47	0,0753	0,0728	0,0702	0,0677	0,0652	0,0627	0,0602	0,0577	0,0552	0,0527	0,0502	0,52
0,48	0,0502	0,0476	0,0451	0,0426	0,0401	0,0376	0,0351	0,0326	0,0301	0,0276	0,0251	0,51
0,49	0,0251	0,0226	0,0201	0,0175	0,0150	0,0125	0,0100	0,0075	0,0050	0,0025	0,0000	0,50
	0,010	0,009	0,008	0,007	0,006	0,005	0,004	0,003	0,002	0,001	0,000	P

Grandes valeurs de u

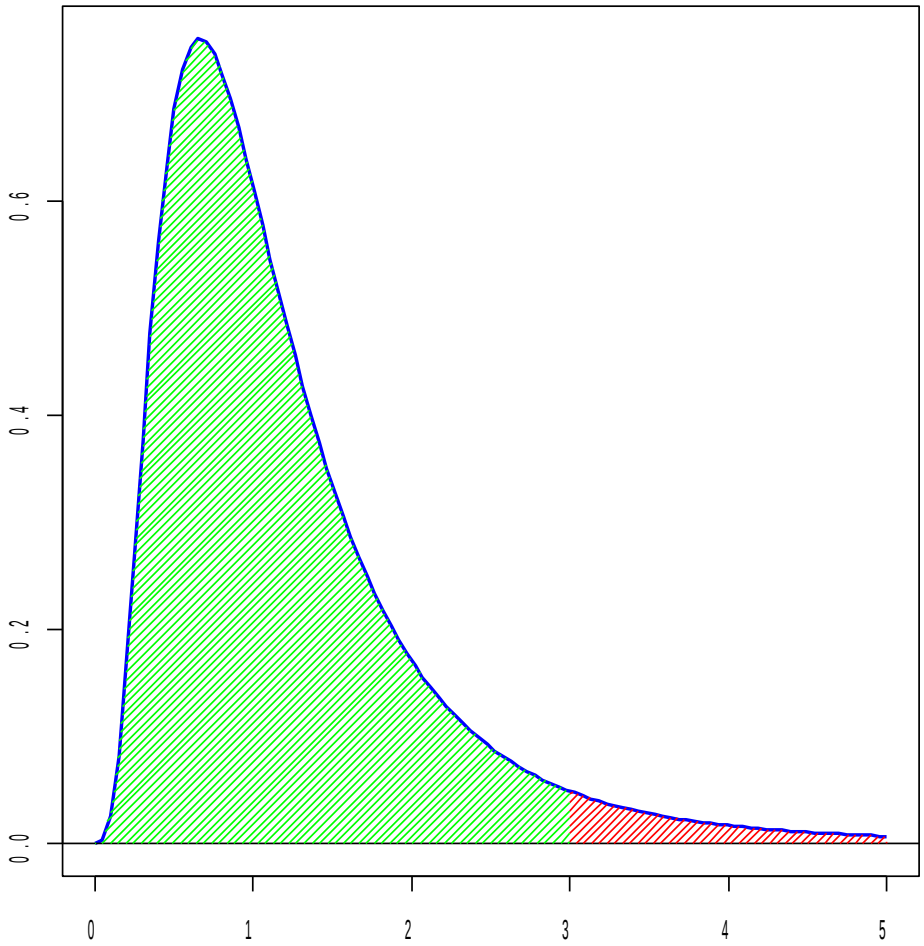
P	0,9999	0,99999	0,999999	0,9999999	0,99999999	0,999999999
u	3,7190	4,2649	4,7534	5,1993	5,6120	5,9978

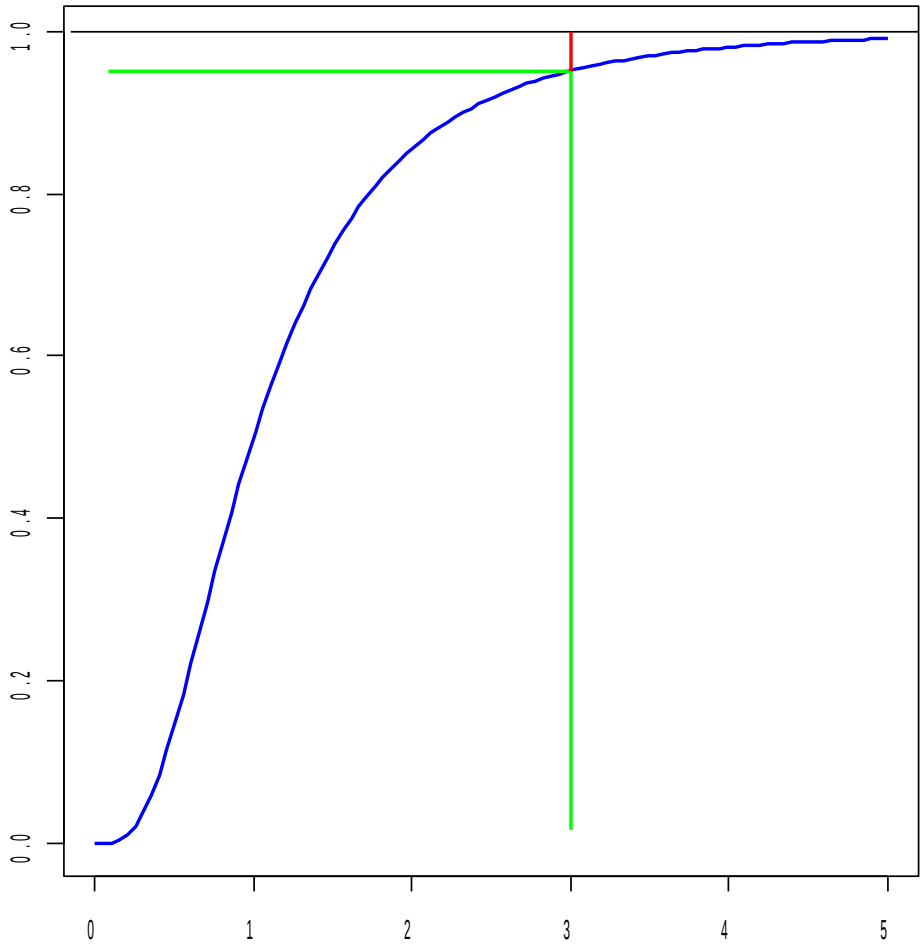
N.B. Si $P < 0.5$, u est négatif.

C

Distribution de probabilité – Fonction de répartition

« aire à droite » = 1 - « aire à gauche »

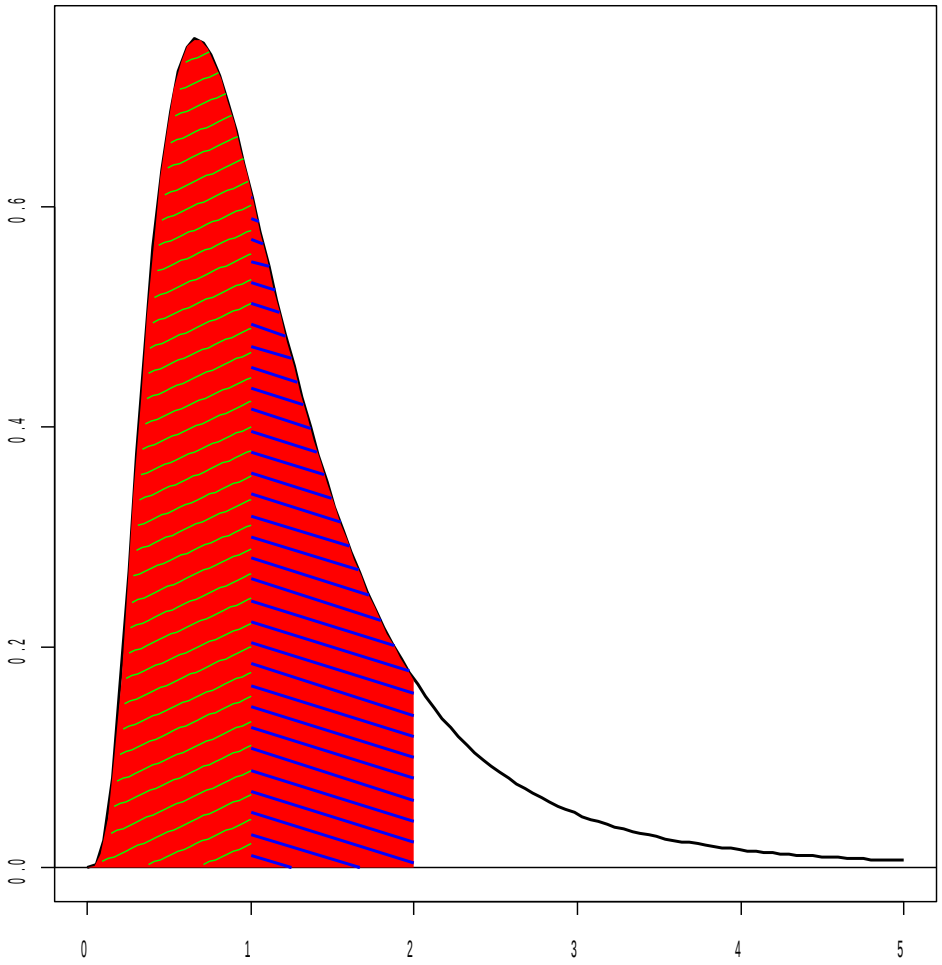


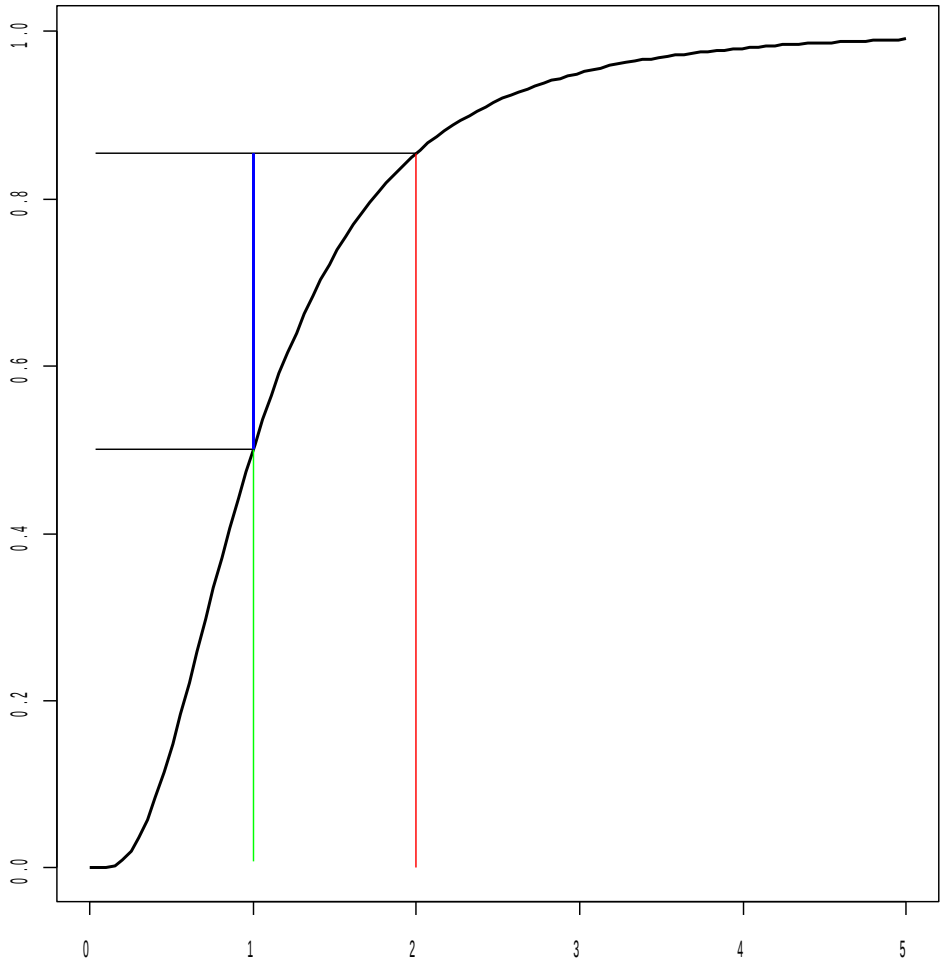


C

Distribution de probabilité – Fonction de répartition

« aire centrale » = « grande aire à gauche » - «petite aire à gauche »





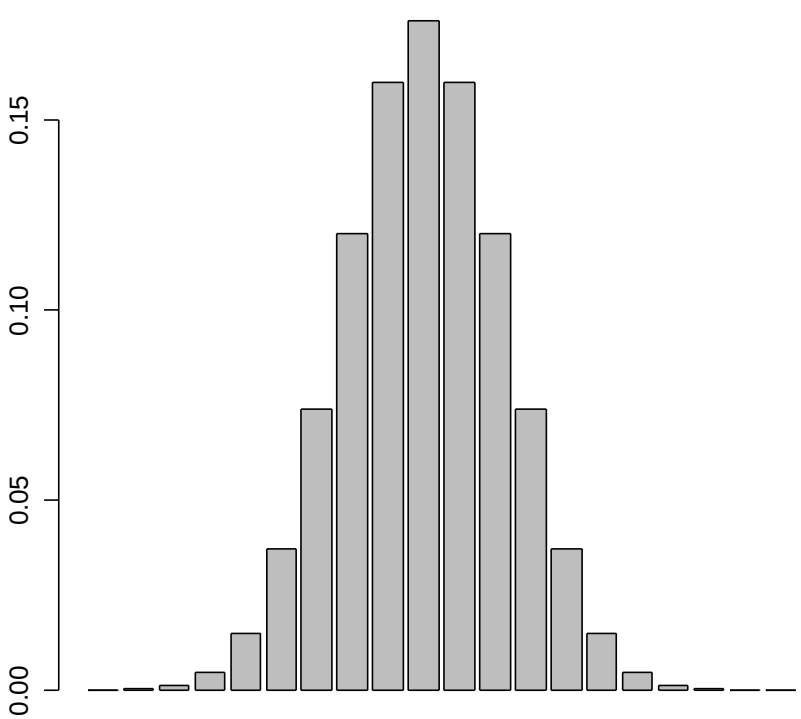
Quelques lois usuelles

Loi binomiale B(n,p)

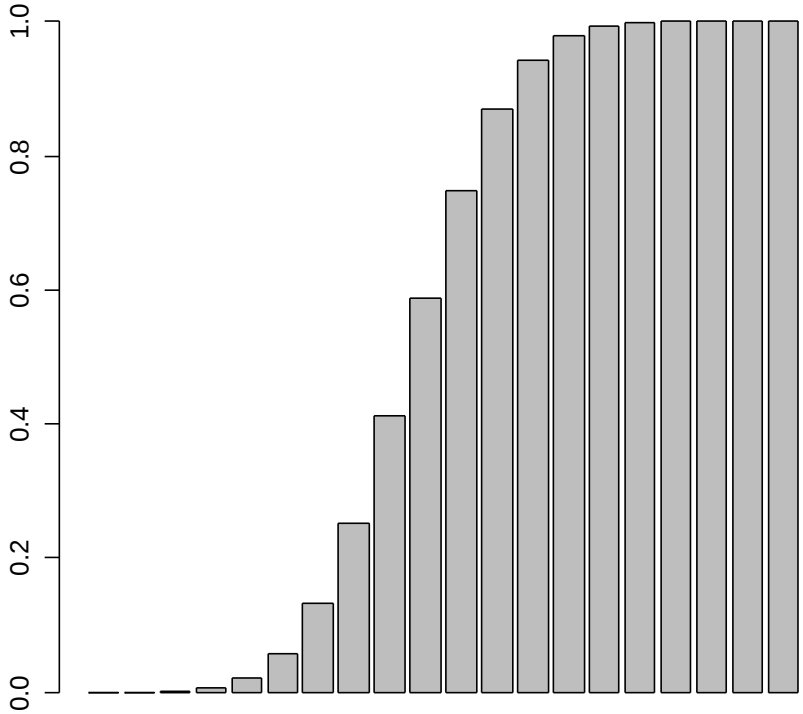
$$P(X=k)=C_n^k\, p^k (1-p)^{n-k}$$

$$C_n^k=\frac{n!}{k!\,(n-k)!}$$

Ex : n=20, p=0.5



Loi de probabilité

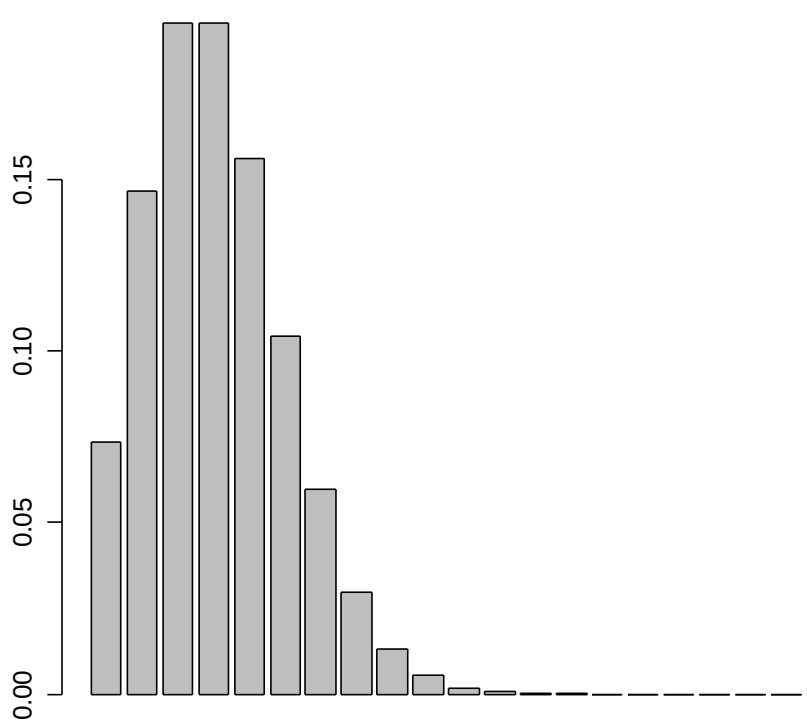


Fonction de répartition

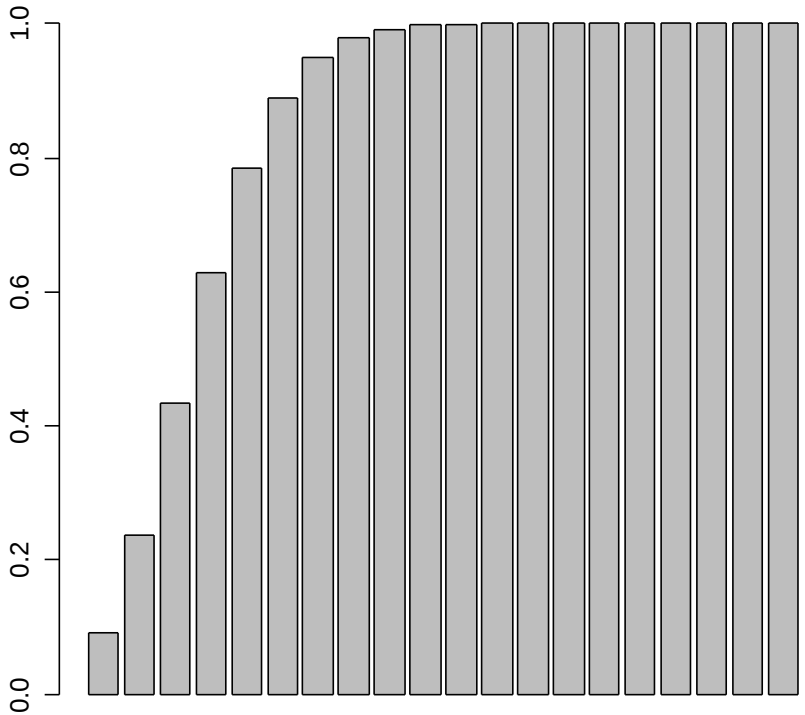
Quelques lois usuelles

Loi de Poisson $P(\lambda)$ $$P(X=k)=e^{-\lambda}\frac{\lambda^k}{k!}$$

Ex : $\lambda=4$



Loi de probabilité



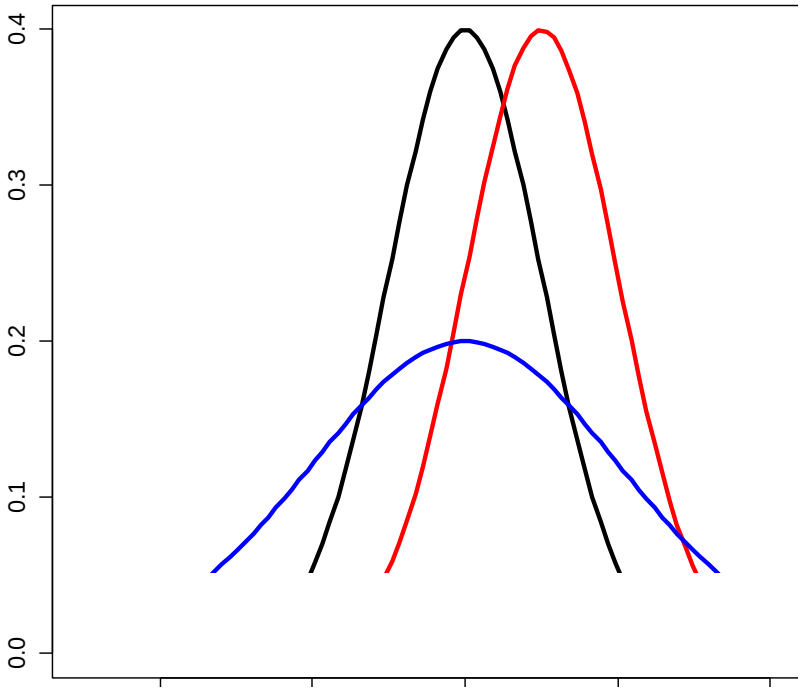
Fonction de répartition

Quelques lois usuelles

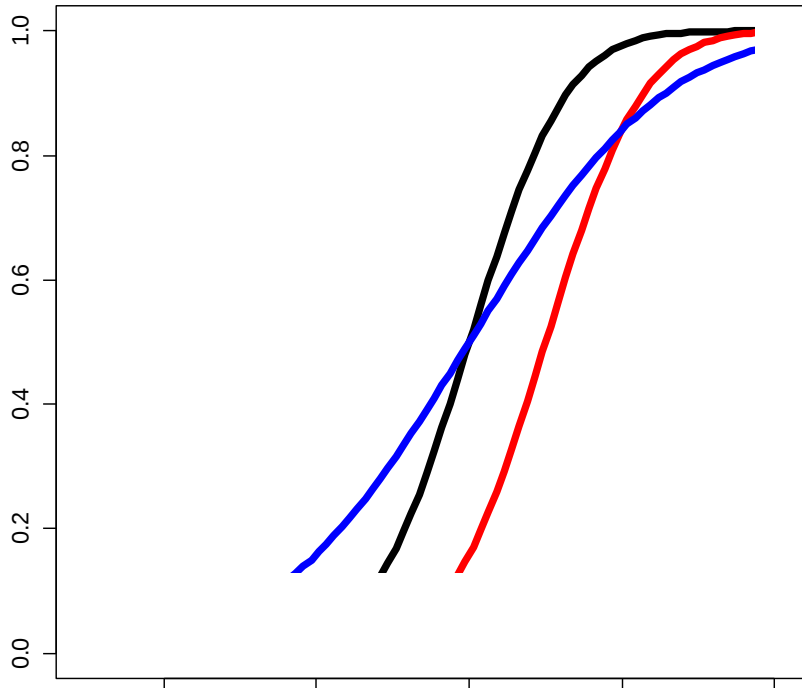
Loi normale $N_{(\mu,\sigma^2)}$

$$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{(x-\mu)^2}{\sigma^2}\right)$$

Ex : $\mathbf{N}_{(0,1)}$ $\mathbf{N}_{(1,1)}$ $\mathbf{N}_{(0,2)}$



Loi de probabilité

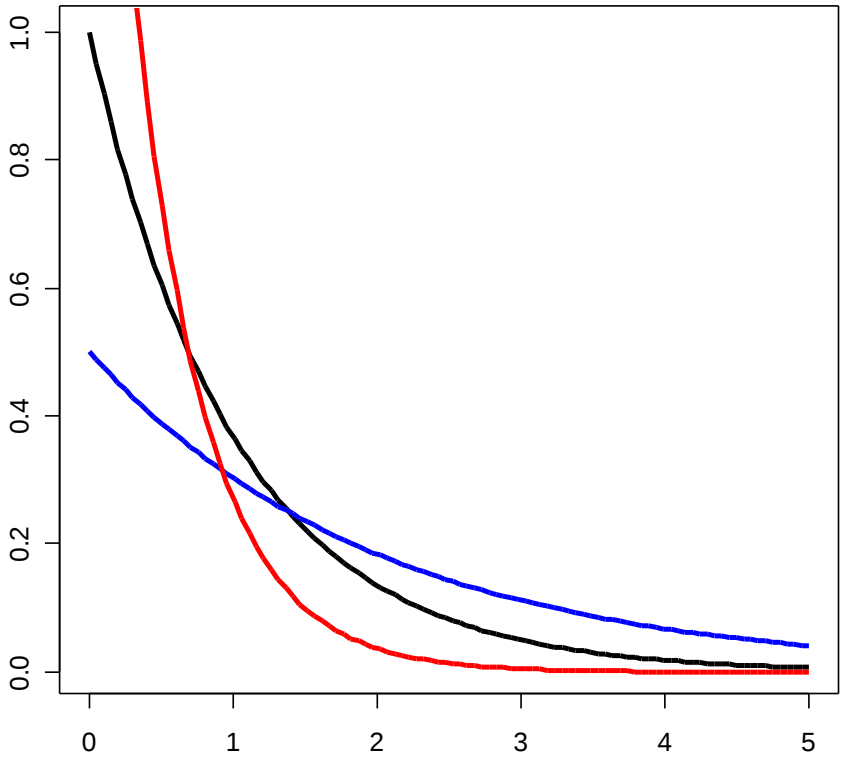


Fonction de répartition

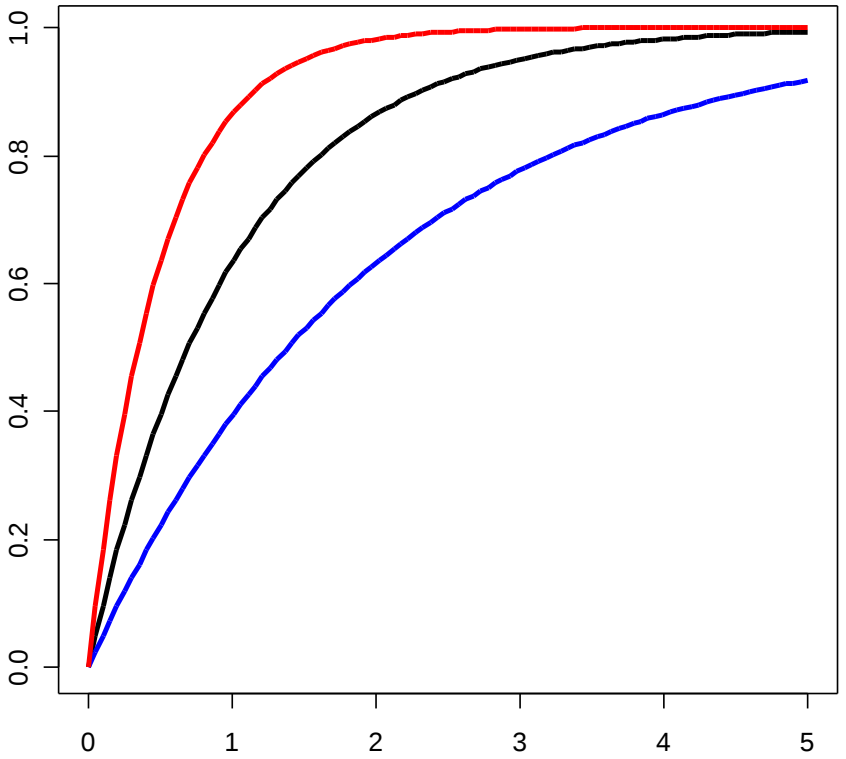
Quelques lois usuelles

Loi exponentielle $E(\lambda)$ $f(x)=\lambda\exp(-\lambda x)$

Ex : λ=1 – 0.5 - 2



Loi de probabilité



Fonction de répartition

Quelques lois usuelles

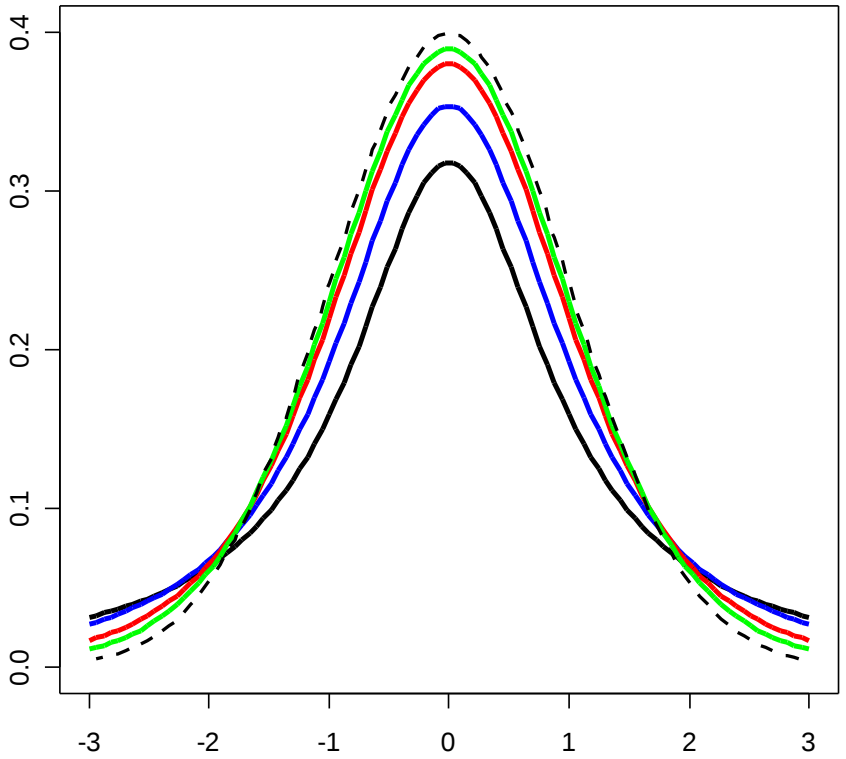
Loi de Student $St(k)$

$$f(x)=\frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\,\Gamma\left(\frac{k}{2}\right)\left(1+\frac{t^2}{k}\right)^{\frac{k+1}{2}}}$$

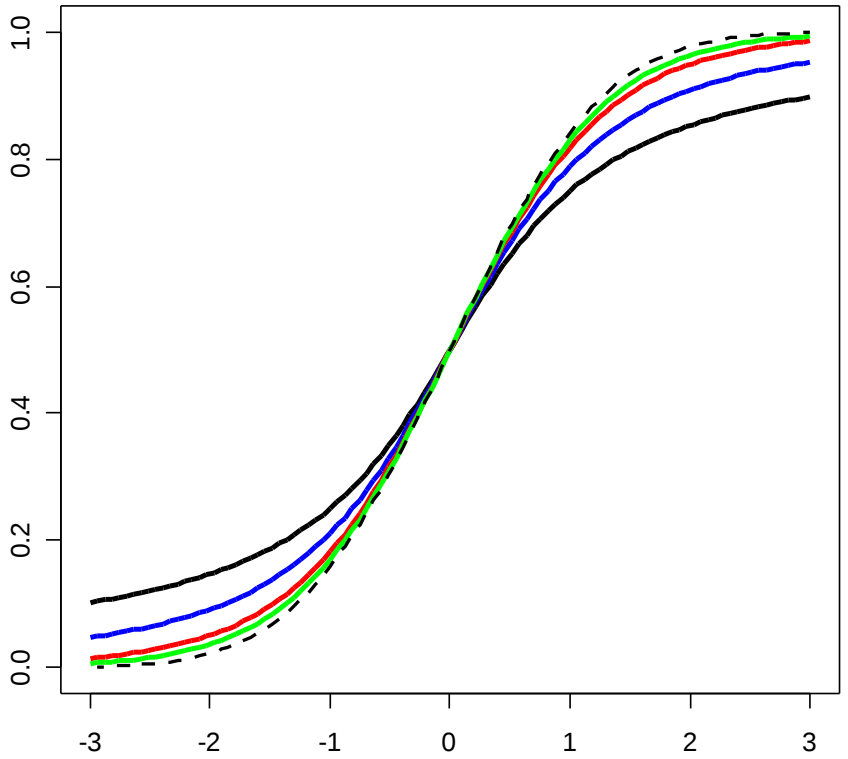
Γ *fonction Gamma d'Euler*

Ex : k=1 - 2 - 5 - 10

Loi normale $\mathbf{N}(0,1)$ **- - -**



Loi de probabilité



Fonction de répartition

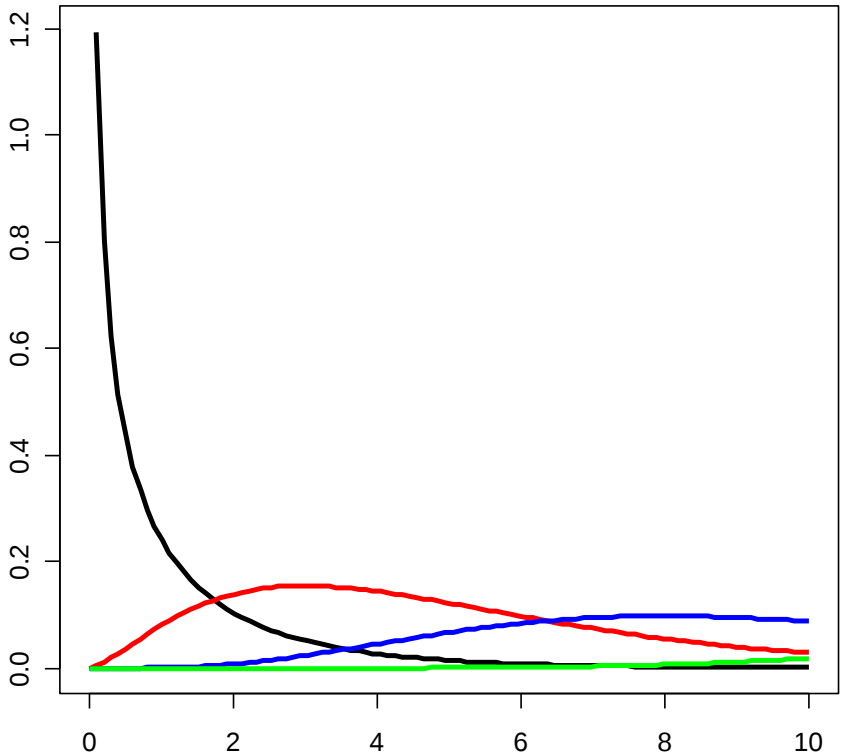
Quelques lois usuelles

Loi du khi-deux $\chi^2(k)$

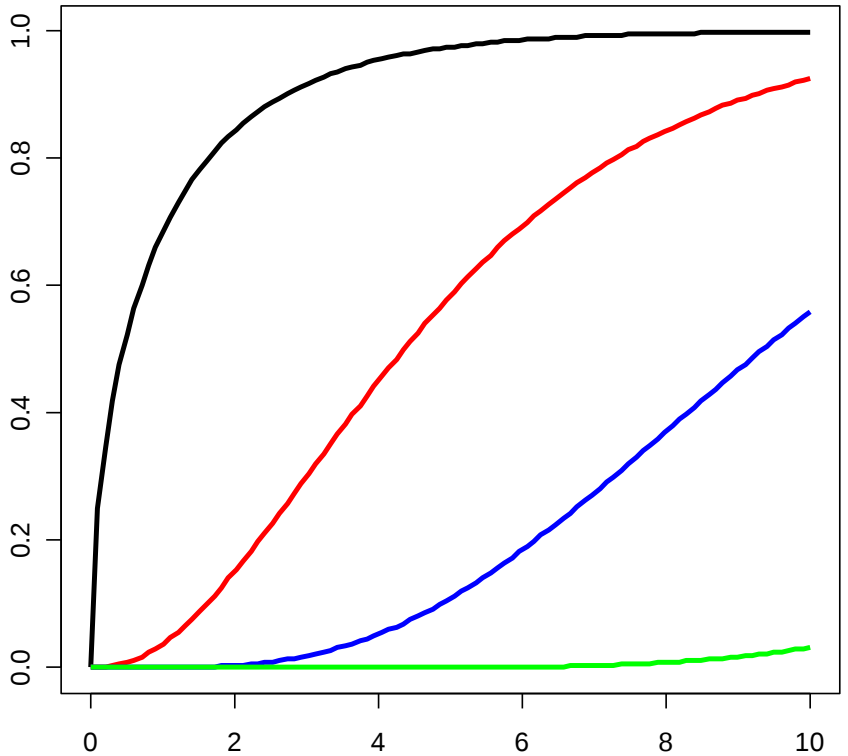
$$f(x)=\frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)}x^{\frac{k}{2}-1}e^{-\frac{x}{2}}$$

Γ *fonction Gamma d'Euler*

Ex : k=1 - 5 - 10 - 20



Loi de probabilité



Fonction de répartition

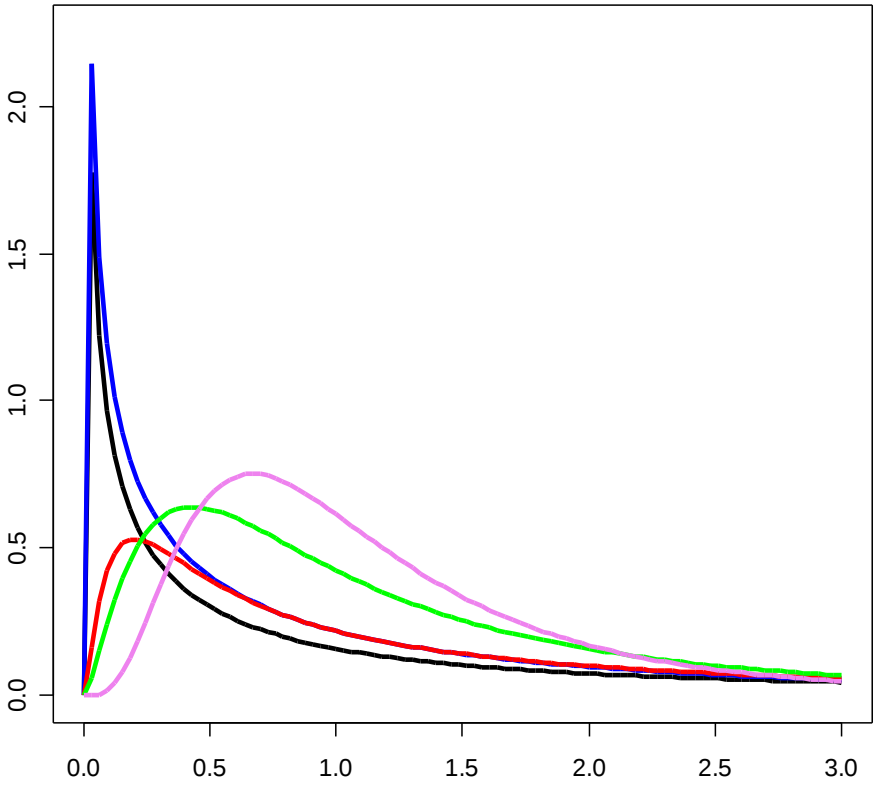
Quelques lois usuelles

Loi de Fisher $F_{(n1,n2)}$

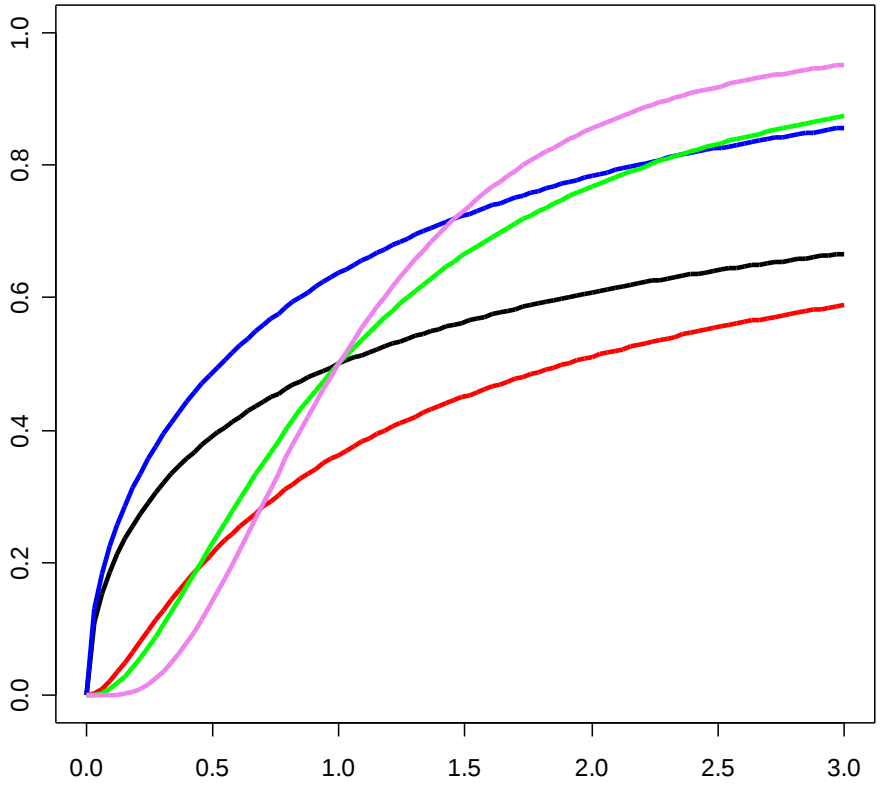
$$f(x)=n_1^{\frac{n1}{2}} n_2^{\frac{n2}{2}} \frac{\Gamma\left(\frac{n_1+n_2}{2}\right) x^{\frac{n_1}{2}-1}}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)(n_1 x+n_2)^{\frac{n_1+n_2}{2}}}$$

Γ *fonction Gamma d'Euler*

Ex : **F**(1,1) **F**(1,5) **F**(5,1) **F**(5,5) **F**(10,10)



Loi de probabilité



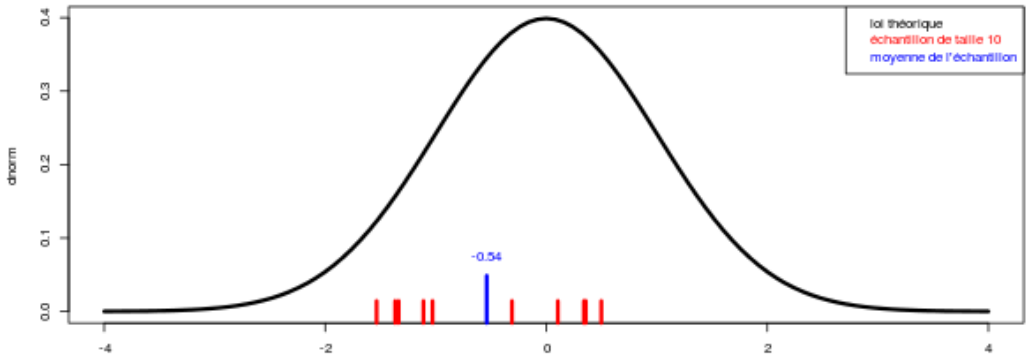
Fonction de répartition

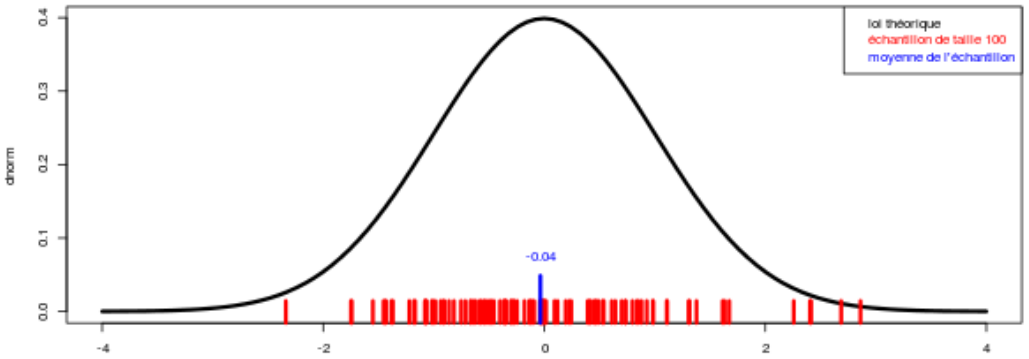
Des simulations

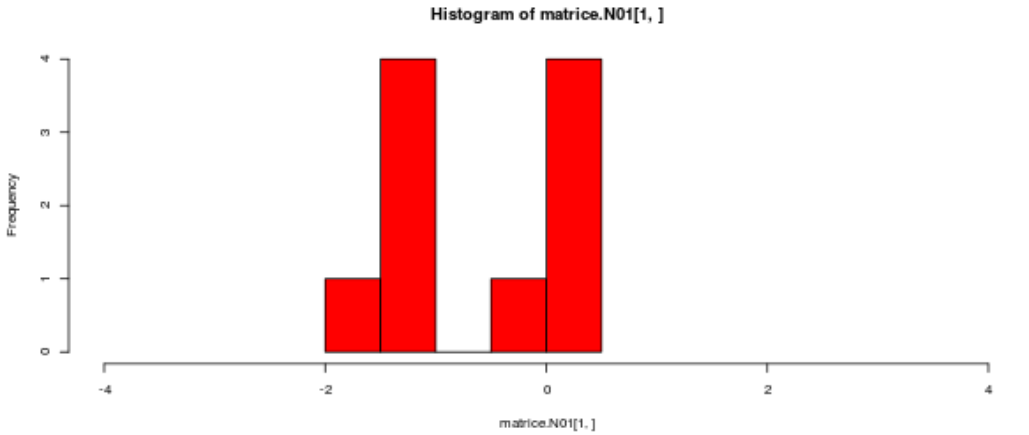
Tirages aléatoires selon une loi normale N(0,1)

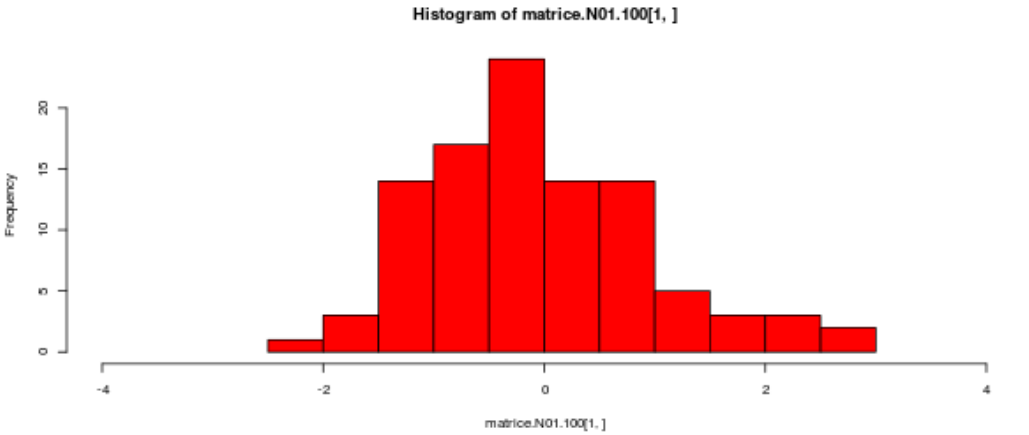
Échantillon de longueur **10**

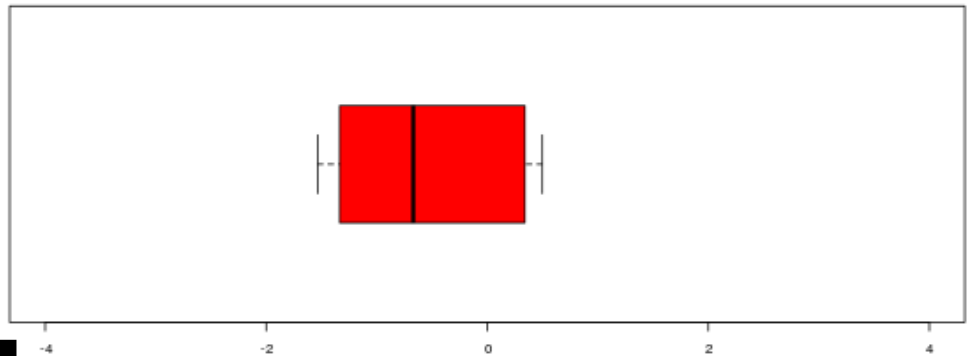
Échantillon de longueur **100**

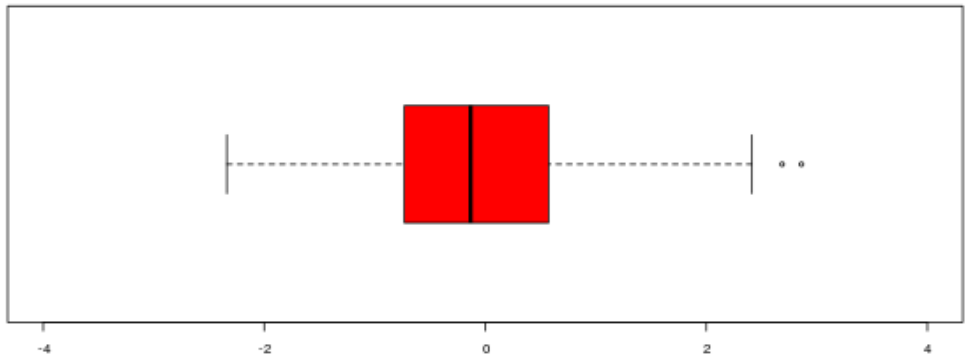








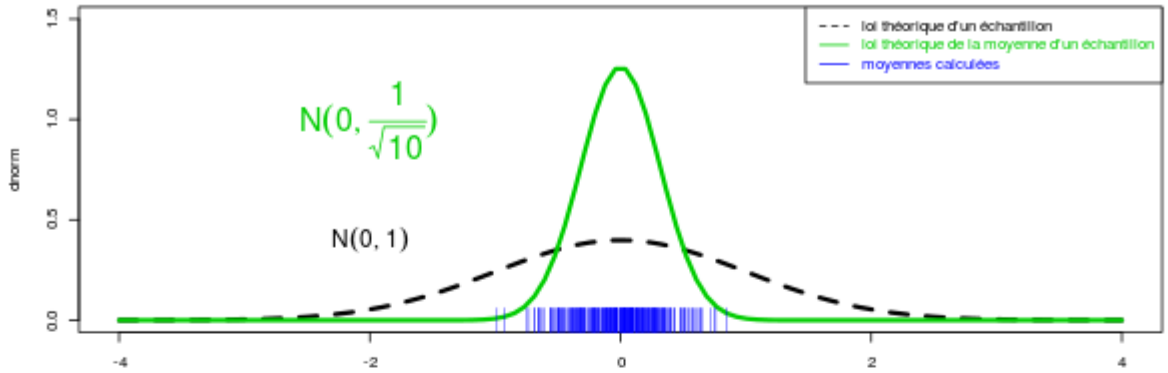


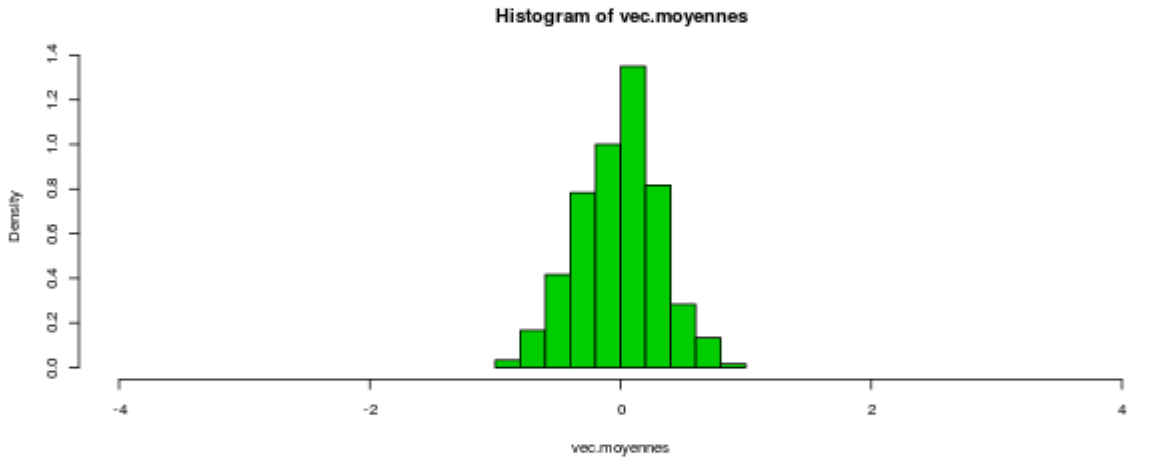


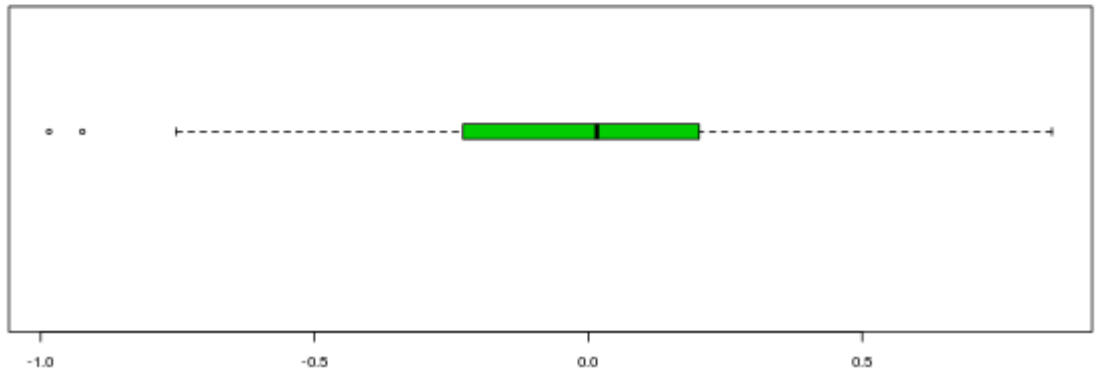
C

Des simulations

Supposons maintenant que nous disposons de **300** échantillons de longueur **10** tous issus d'une loi normale N(0,1). On peut calculer la moyenne de chaque échantillon et avoir ainsi un vecteur de moyennes de longueur 300.







- La moyenne des 300 moyennes vaut (sur l'illustration) :

-0.017

- L'écart-type des 300 moyennes vaut (sur l'illustration) :

0.303

ce qui proche de la valeur théorique donnée par $\frac{1}{\sqrt{10}} \approx 0.316$

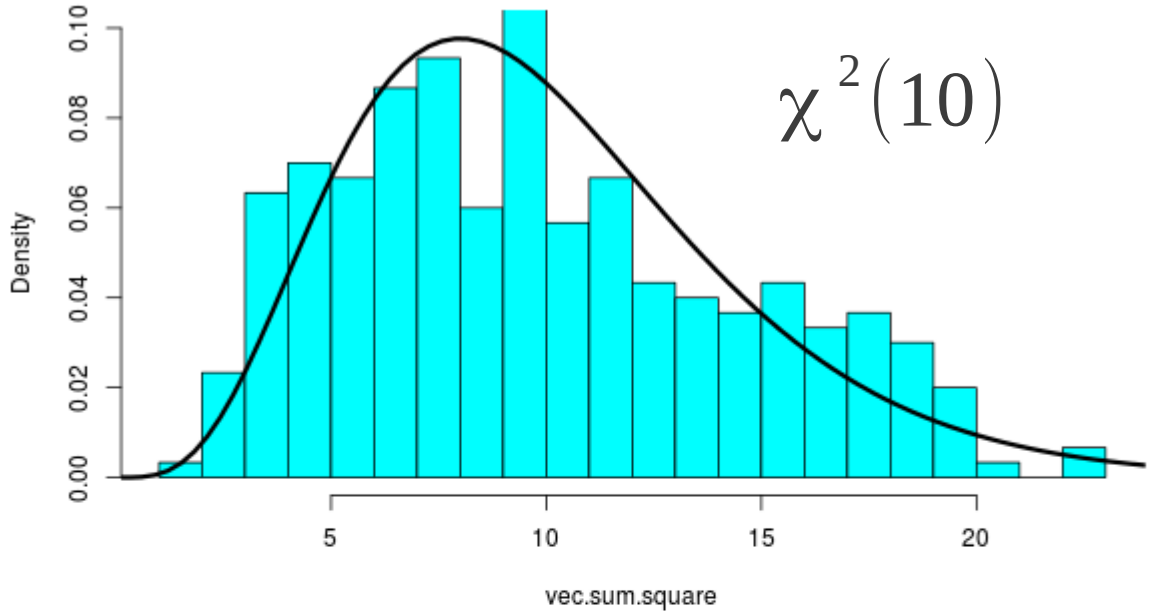
Théorie

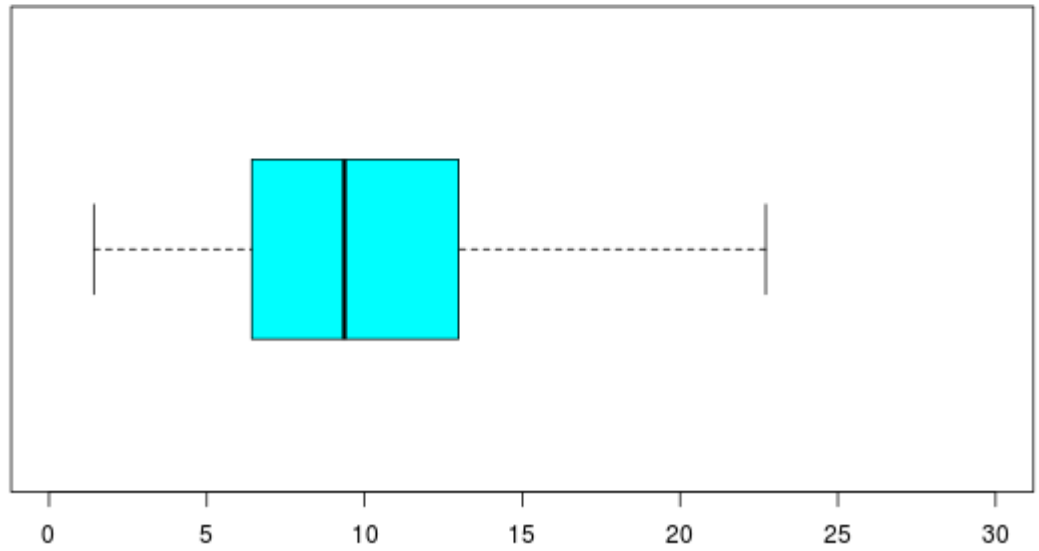
$$\begin{cases} X_i \sim N(\mu, \sigma^2), i = 1, \ldots, n \\ \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \end{cases}$$

C

Des simulations

Intéressons-nous maintenant à la somme des carrés des **10** valeurs de chaque échantillon. Nous disposons d'un vecteur de longueur **300** dont l'histogramme est représenté ci-contre. La courbe représente la loi théorique de cette distribution c'est à dire une loi de χ^2 à 10 degrés de liberté.





Théorie

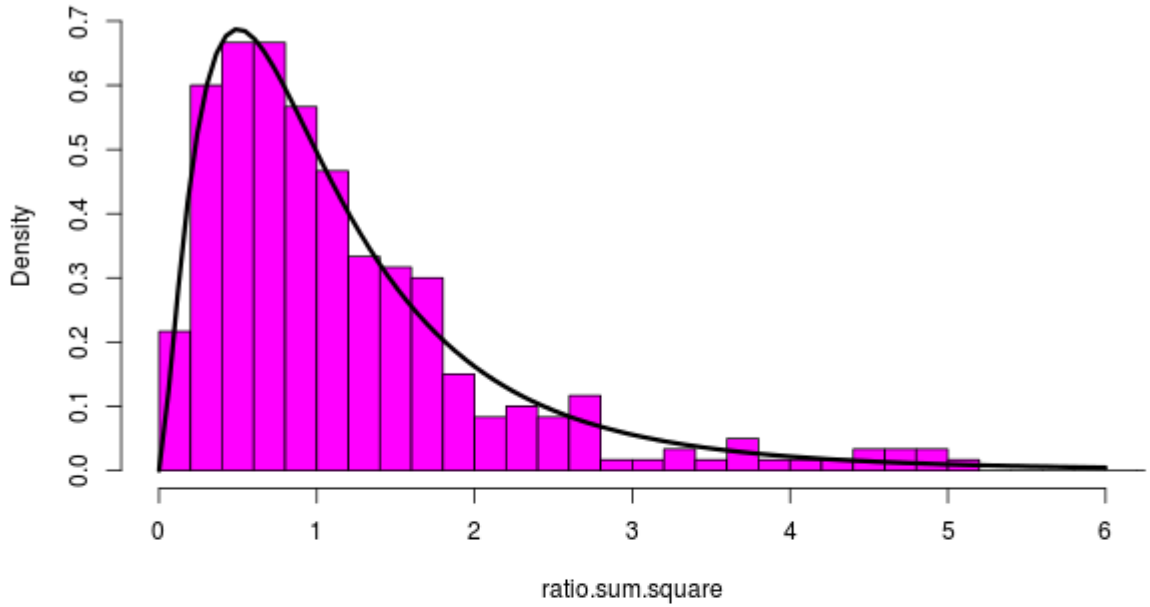
$$X_i \sim N(\mu, \sigma^2), i = 1, \ldots, n$$

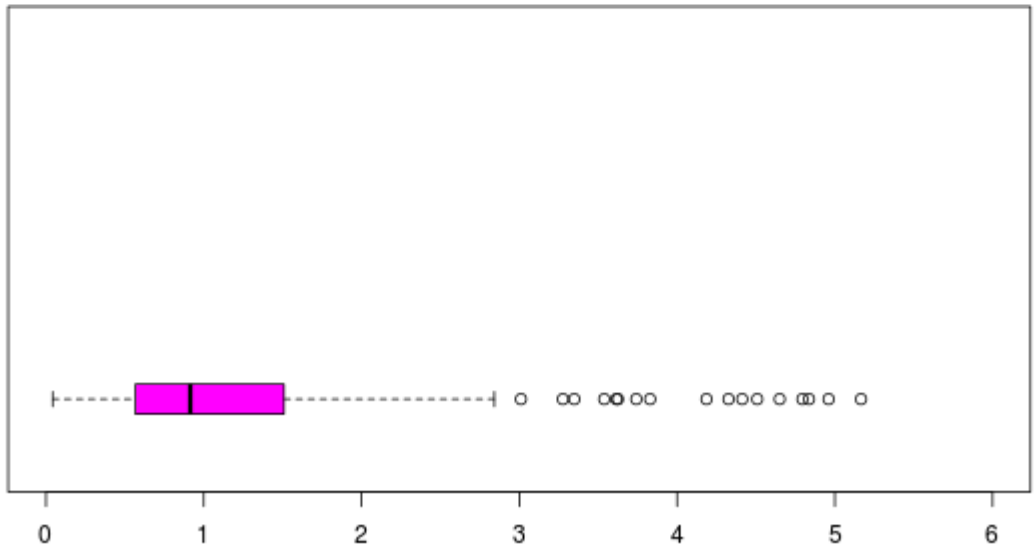
$$\sum_{i=1}^{n} X_i^2 \sim \chi^2(n)$$

C

Des simulations

Supposons maintenant que nous disposons de **300** échantillons de longueur **10** et de **300** échantillons de longueur **5** tous issus d'une loi normale N(0,1).
On sait que les sommes de carrés des éléments des vecteurs de longueur 10 et 5 suivent des lois de χ^2 à 5 et 10 degrés de liberté.
Le rapport des 2 variables « sommes de carrés » divisées par la taille des échantillons suit alors une loi de Fisher à 5 et 10 degrés de liberté.





Théorie

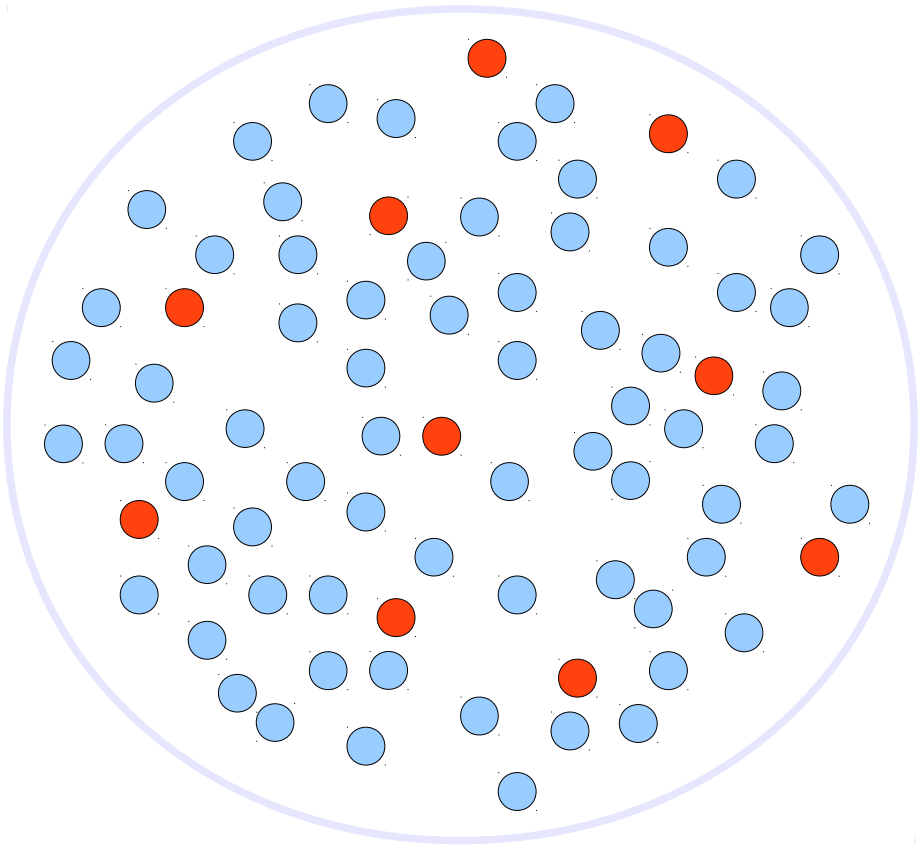
$$C1 \sim \chi^2(n_1); C2 \sim \chi^2(n_2)$$

$$\frac{C_1/n_1}{C_2/n_2} \sim F(n_1, n_2)$$

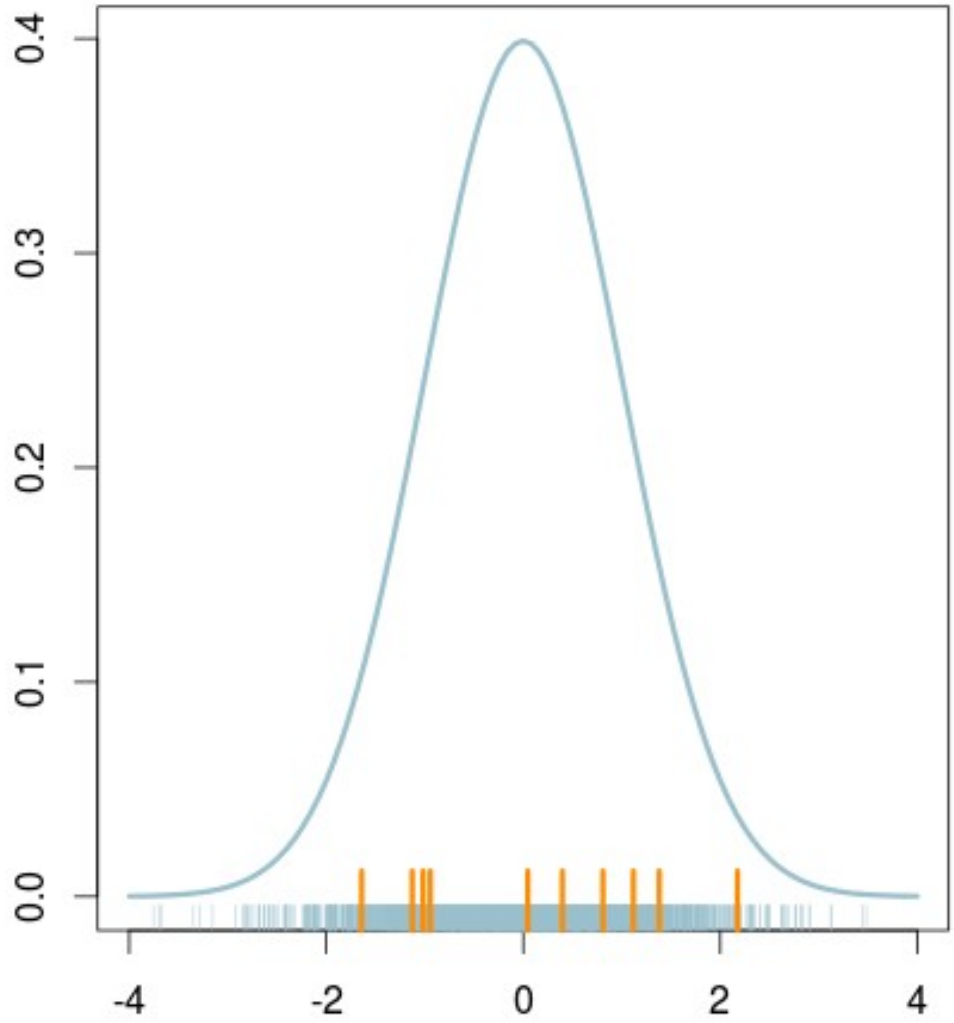
Statistique inférentielle

- Tirer des conclusions à l'échelle d'une **population** à partir d'informations recueillies sur un **échantillon**.
- Sondage, recensement, échantillon représentatif...
- Lorsque l'on avance des informations quantitatives à l'échelle de la population, on ne parle plus de mesure mais d'**estimation**.
- Les mesures effectuées sur l'échantillon sont des **observations** de la variable aléatoire traduisant le phénomène à l'échelle de la population.

Statistique inférentielle



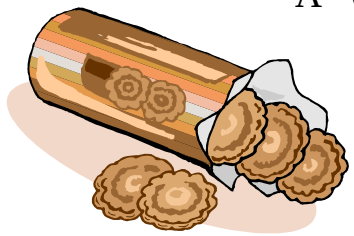
Population
Échantillon



Test statistique : ce qui est observé sur un échantillon permet-il d'invalider une hypothèse faite sur la population ?

D

Estimation



Dans un fabrique de biscuit, le procédé mis en œuvre pour vérifier l'aspect moelleux du produit fini consiste à plier le biscuit et à mesurer l'angle d'inclinaison nécessaire pour le casser (un tel test est dit destructif). La règle étant qu'un bon biscuit doit avoir un angle de rupture de 50° (valeur fictive) : si l'angle est inférieur, le biscuit est trop sec, s'il est supérieur, le biscuit est trop moelleux. Tout lot de biscuit doit être validé avant d'être commercialisé.

Il va de soi qu'un biscuit cassé n'est pas commercialisable ainsi qu'un biscuit n'étant pas convenablement moelleux (angle de rupture $\neq 50°$).

Dans de telles conditions, **il est impossible de tester l'ensemble des biscuits** (test destructif). Il est donc nécessaire d'effectuer les mesures sur un **échantillon représentatif** de la population des biscuits (éviter par exemple de prendre les n premiers ou les n derniers biscuits fabriqués dans une journée ou, sur une même ligne de production si plusieurs fonctionnent en parallèle). L'angle moyen de rupture calculé sur l'échantillon est un **estimateur** de cet angle chez les biscuits du même lot (aux conditions de fabrication analogues).

Test statistique

Exemple : Fabrication industrielle de biscuits dont l'angle de rupture doit être de 50°. Des facteurs incontrôlés font que cet angle est aléatoire.

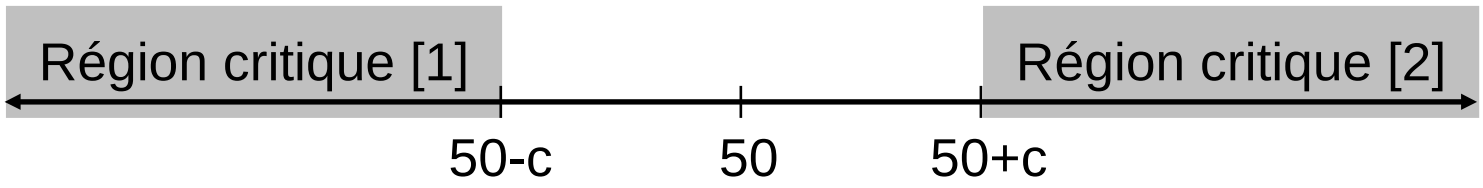
Question : comment décider qu'un lot est conforme ?

Hypothèses :

H0 : le lot est conforme ($\mu=50$)
H1 : le lot n'est pas conforme ($\mu\neq50$)

Pour trancher entre les 2 hypothèses, on tire au hasard un échantillon de *n* biscuits et on en mesure l'angle de rupture $(X_i)_{i=1,...n}$. Chaque X_i suit une loi $N(\mu,\sigma^2)$.

Règle de décision (principe): Rejet de H0 $\Leftrightarrow \overline{X} \notin [50\text{-}c\ ;\ 50+c]$



Test statistique : risques d'erreur

Rappel :
H0 : lot conforme
H1 : lot non conforme

		Décision	
		H1 (rejet de H0)	H0 (accept. H0)
Réalité	H0	α	Bonne décision
	H1	Bonne décision	β

Interprétation des risques (en termes de biscuit) :

- α : rejeter le lot de biscuits alors qu'il est conforme (gaspillage !)

  ➔ *Le patron ne va pas être content.*

- β : déclarer conforme (et donc vendre) des biscuits « défectueux »

  ➔ *Dans ce cas, c'est le client qui n'est pas content.*

Région critique et risque α

Règle de décision : Rejet de H0 $\Leftrightarrow \bar{X} \notin$ [50-c ; 50+c]

α = P[Rejeter H0 // H0 vraie] = P[$\bar{X} \notin$ [50-c ; 50+c] // μ =50]

Sous H0 (μ =50)

$X_i \sim N(\mu,\sigma^2)$ ➔ $\bar{X} \sim N(\mu,\sigma^2/n)$

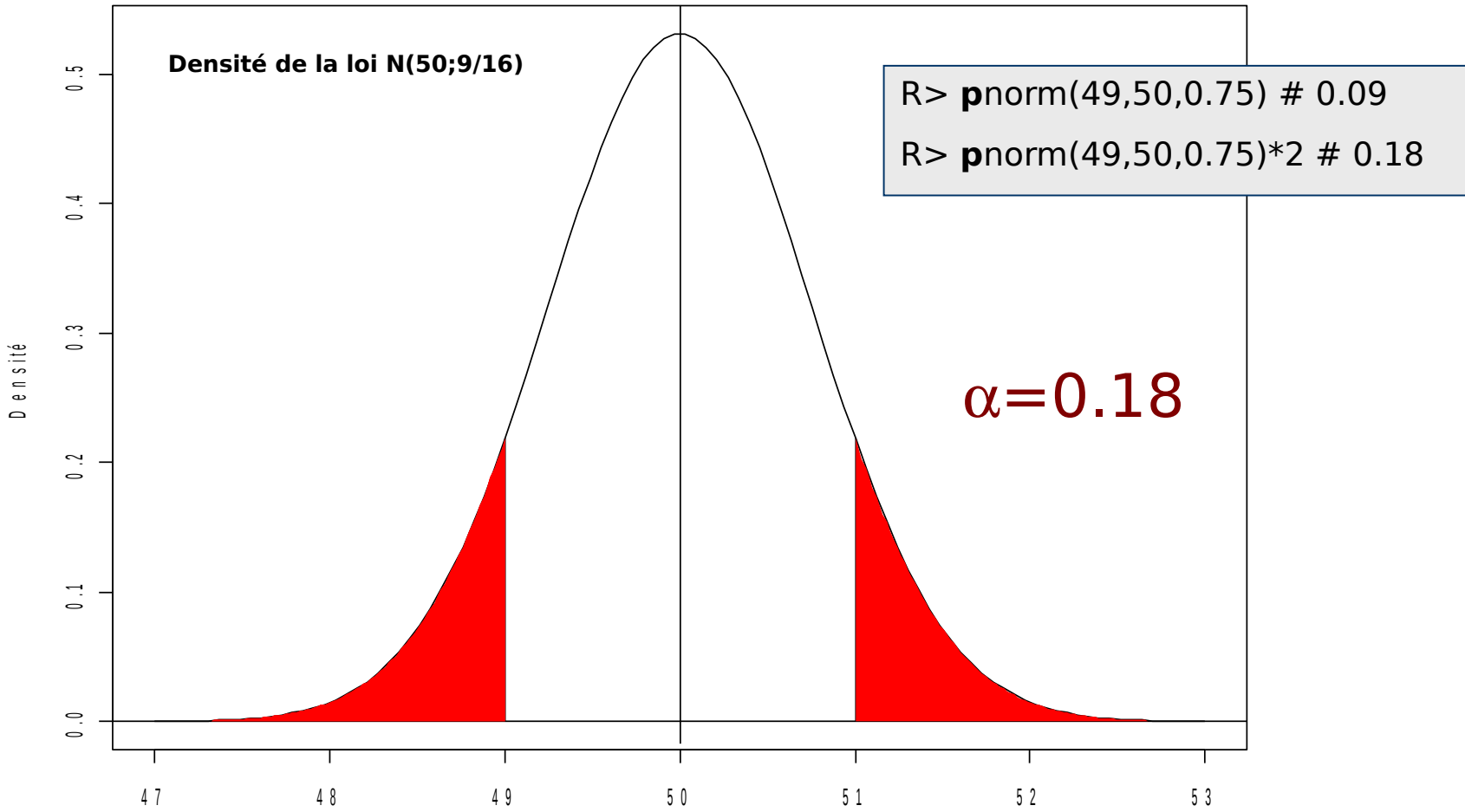
Pour l'application numérique : n=16 et σ^2=9

α=P[« N(50,9/16) » $\notin$ [50-c ; 50+c]]

Le risque α est la probabilité qu'une variable aléatoire suivant une loi normale de moyenne 50 et de variance 9/16 n'appartienne pas à l'intervalle [50-c ; 50+c].

Région critique et risque α

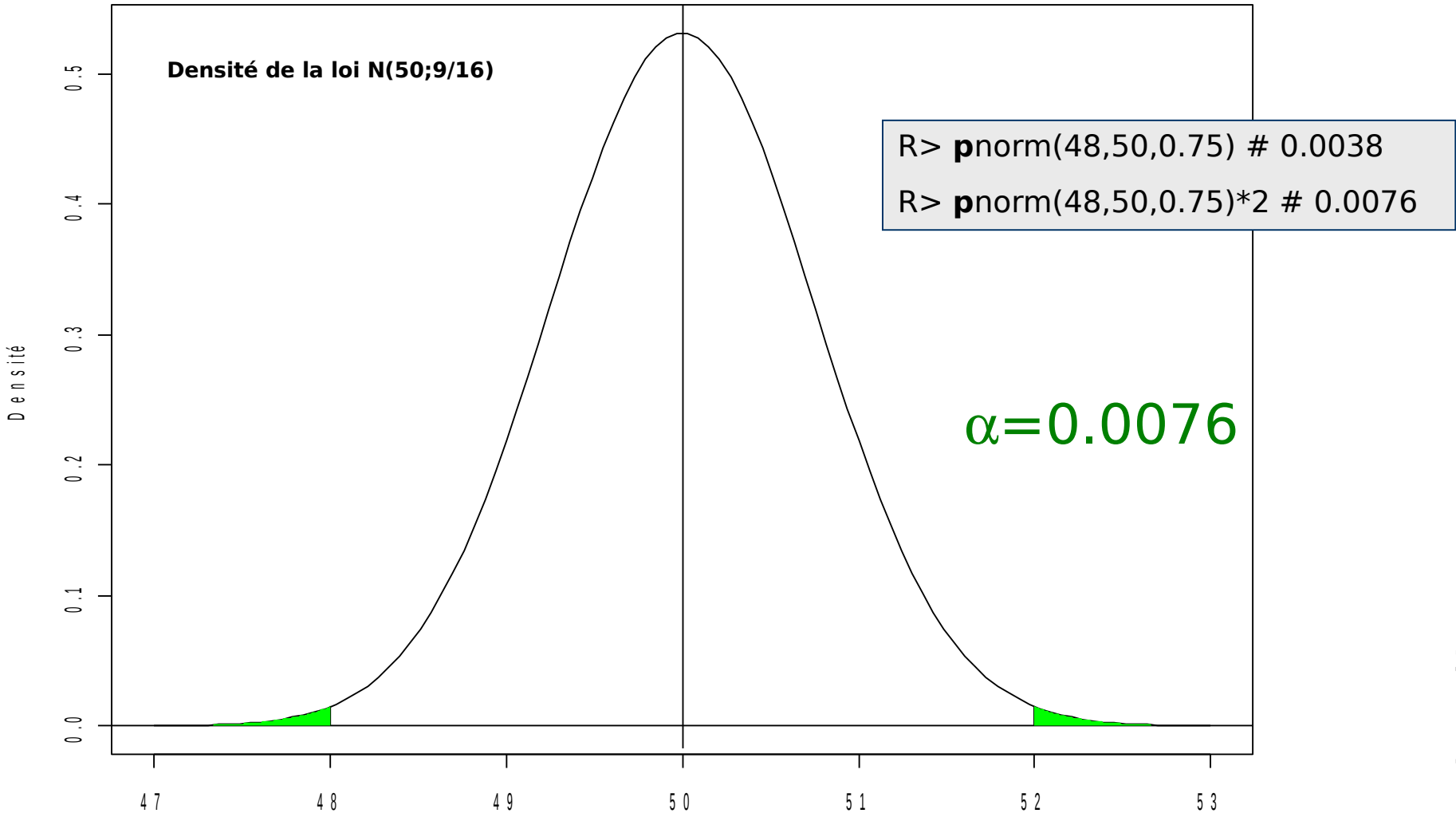
➔ Exemple : prenons **c=1**, la région critique est **![49 ; 51]**. Calculons le risque α associé. $\alpha = P[\text{« } N(50,9/16) \text{ »} \notin [49 ; 51]$



D **![49 ; 51] =]-∞ ; 49] ∪ [51 ; +∞ [**

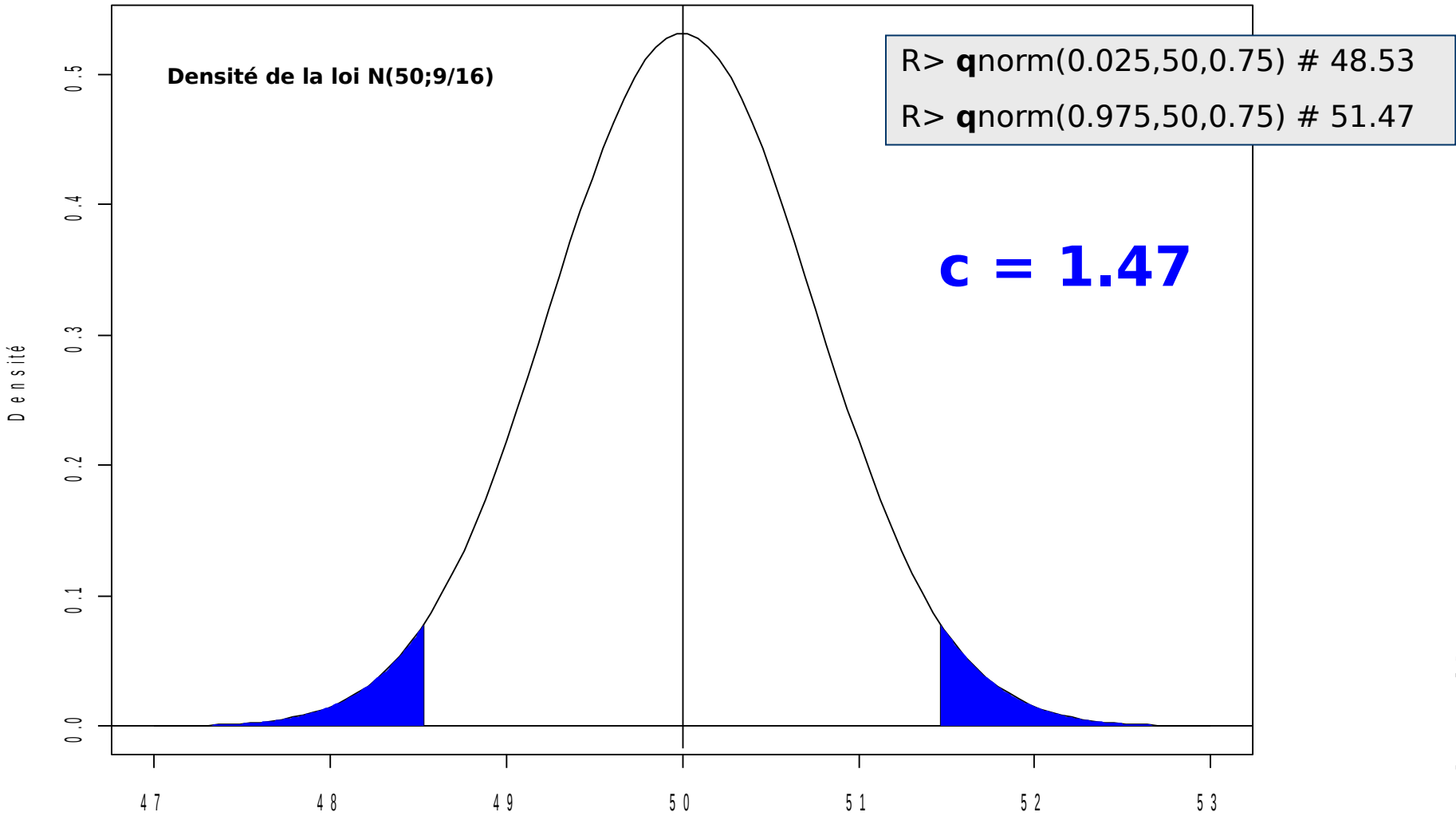
Région critique et risque α

➔ Exemple : prenons **c=2**, la région critique est **![48 ; 52]**. Calculons le risque α associé. $\alpha = P[\text{« } N(50,9/16) \text{ »} \notin [48 ; 52]]$



Région critique et risque α

➔ Trouver c tel que : $\alpha = P[\text{« } N(50,9/16) \text{ »} \notin [50-c\ ;\ 50+c]\ ] = \mathbf{0.05}$



P-value

- « Degré de significativité »
- C'est la plus petite des valeurs de α pour lesquelles la valeur observée de la statistique de test conduit au rejet de H0.
- C'est donc la probabilité d'obtenir, sous H0, la valeur observée de la statistique de test ou une valeur plus extrême.
- Il est équivalent de :
  - Comparer la statistique de test avec la valeur limite de la région critique (calculée à partir de α).
  - Comparer la p-value et α : rejet de H0 $\Leftrightarrow$ p-value < α

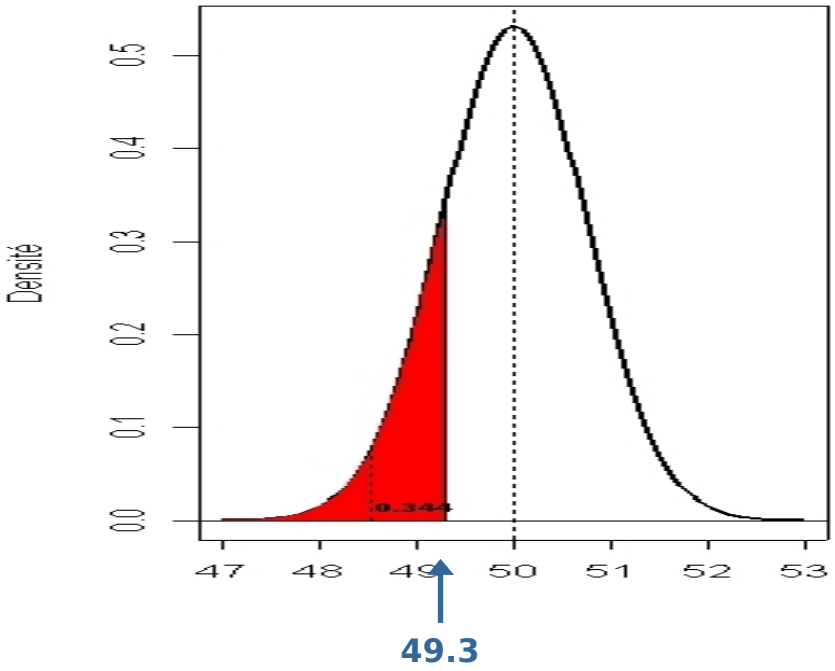
P-value (exemple)

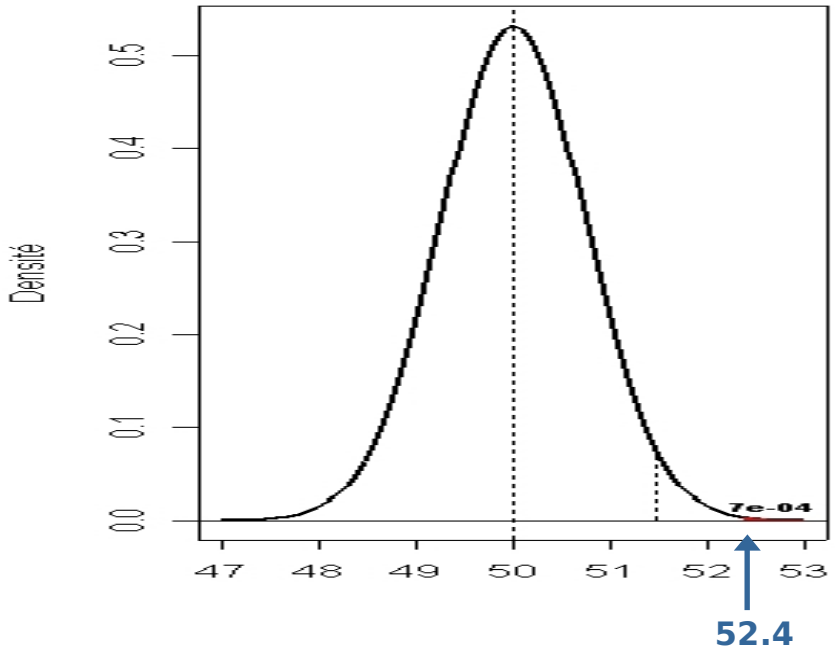
Cas 1) : Jour J1, 16 biscuits tirés au hasard, angle moyen : **49.3**

Cette valeur **n'est pas dans la région critique** (![48.53 ; 51.47]), on ne peut pas rejeter H0, la production du jour est probablement conforme. La p-value associée à la valeur 49.3 est environ 0.17 ce qui est supérieur au seuil de 5%.

- **Cas 2)** : Jour J2, 16 biscuits tirés au hasard, angle moyen : **52.4**

Cette valeur **est dans la région critique**, on rejette H0, la production du jour n'est pas conforme (au seuil de 5%). La p-value associée à la valeur 52.4 est de l'ordre de 0.0007 ce qui est inférieur au seuil de 5%.





Et le risque β ?

$$\beta = P[\text{Accepter H0 // H1 vraie}] = P[\bar{X} \in [50-c\ ;\ 50+c] \ //\ \mu = \mathbf{???}]$$

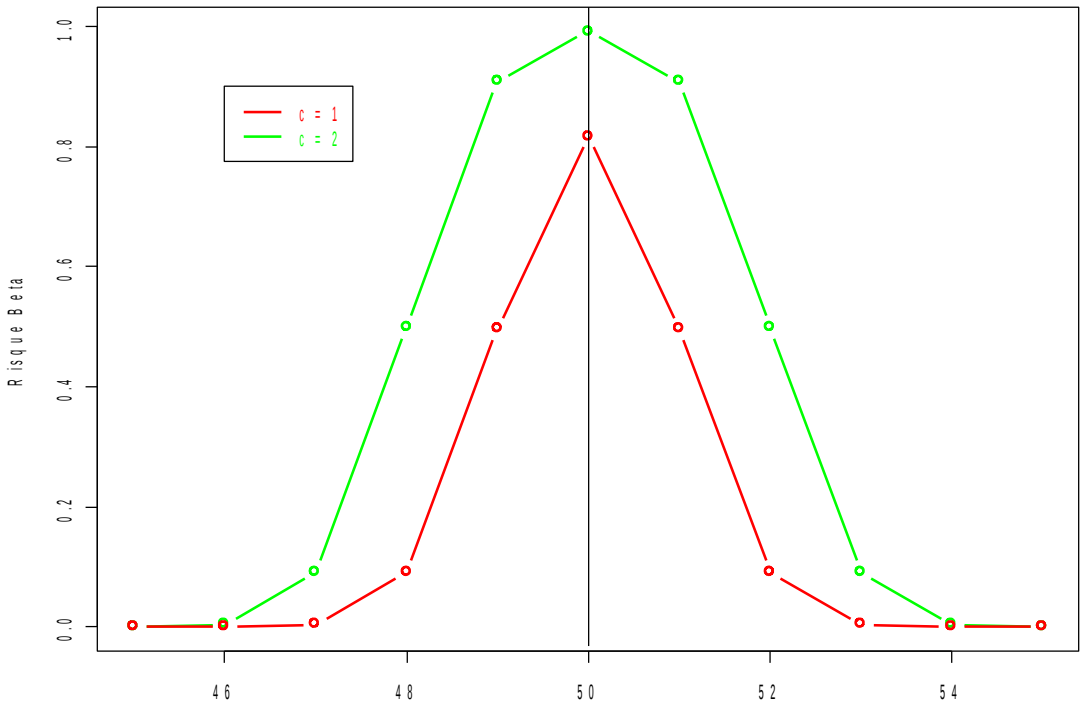
Le calcul explicite du risque β nécessite des valeurs de μ.

$$\beta(\mu) = P[\text{«}N(\mu, 9/16)\text{»} \in [50-c;50+c]\ ]$$

Pour **c=1 :** RC=!**[49 ; 51]** - α=0.18

R> **p**norm(51,45,0.75)-**p**norm(49,45,0.75)

β(**45**)=4.10-8
β(**46**)=3.10-5
β(**47**)=0.0038
β(**48**)=0.091
β(**49**)=0.496
β(**50**)=0.818
β(**51**)=0.496
β(**52**)=0.091
β(**53**)=0.0038
β(**54**)=3.10-5
β(**55**)=4.10-8



Pour **c=2 :** RC=!**[48 ; 52]** - α=0.0076

R> **p**norm(52,45,0.75)-**p**norm(48,45,0.75)

β(**45**)=3.10-5
β(**46**)=0.0038
β(**47**)=0.091
β(**48**)=0.5
β(**49**)=0.91
β(**50**)=0.99
β(**51**)=0.91
β(**52**)=0.5
β(**53**)=0.091
β(**54**)=0.0038
β(**55**)=3.10-5

Puissance d'un test

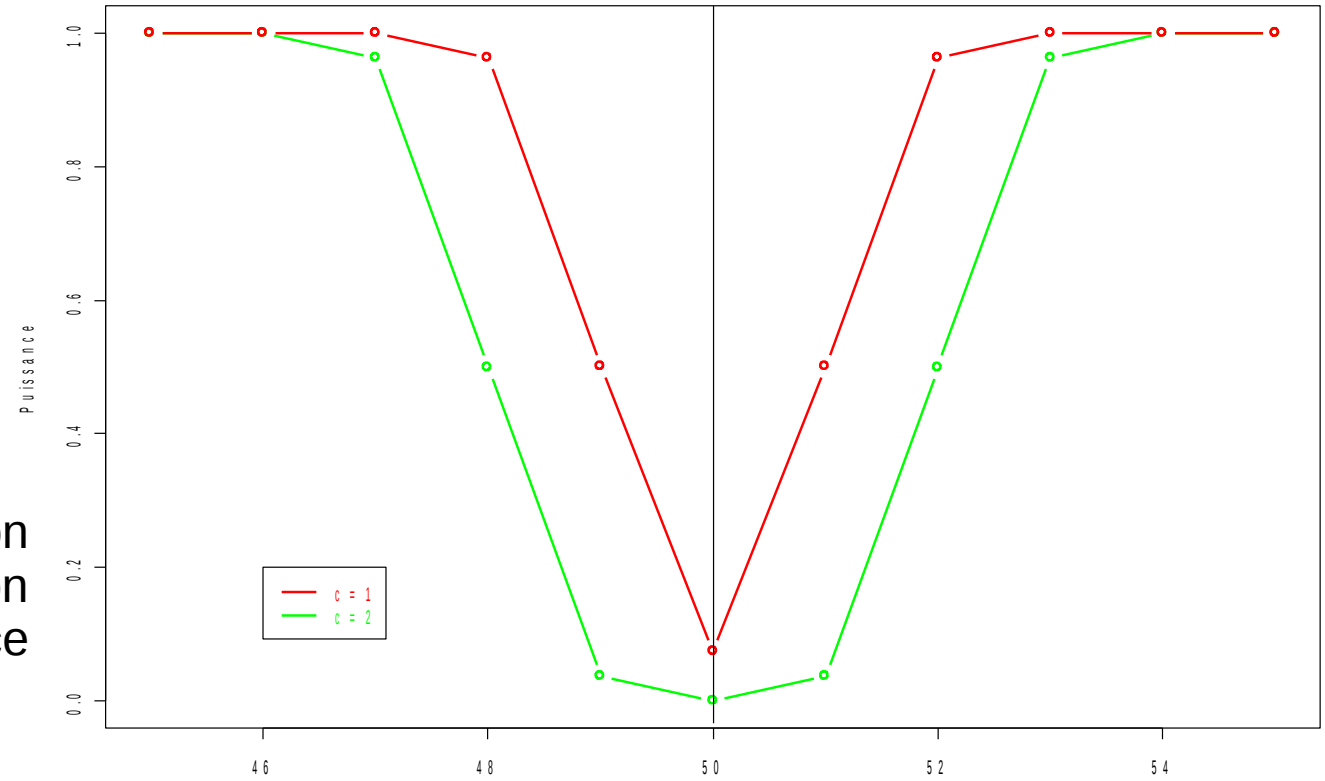
La puissance d'un test est la probabilité de détecter une différence (rejeter H0) lorsqu'elle existe.

$$\mathbf{P} = 1-\beta$$

$$= 1-P[\text{Accepter H0 // H1 vraie}]$$

$$= P[\text{Rejeter H0 // H1 vraie}]$$

Représentation de la fonction puissance



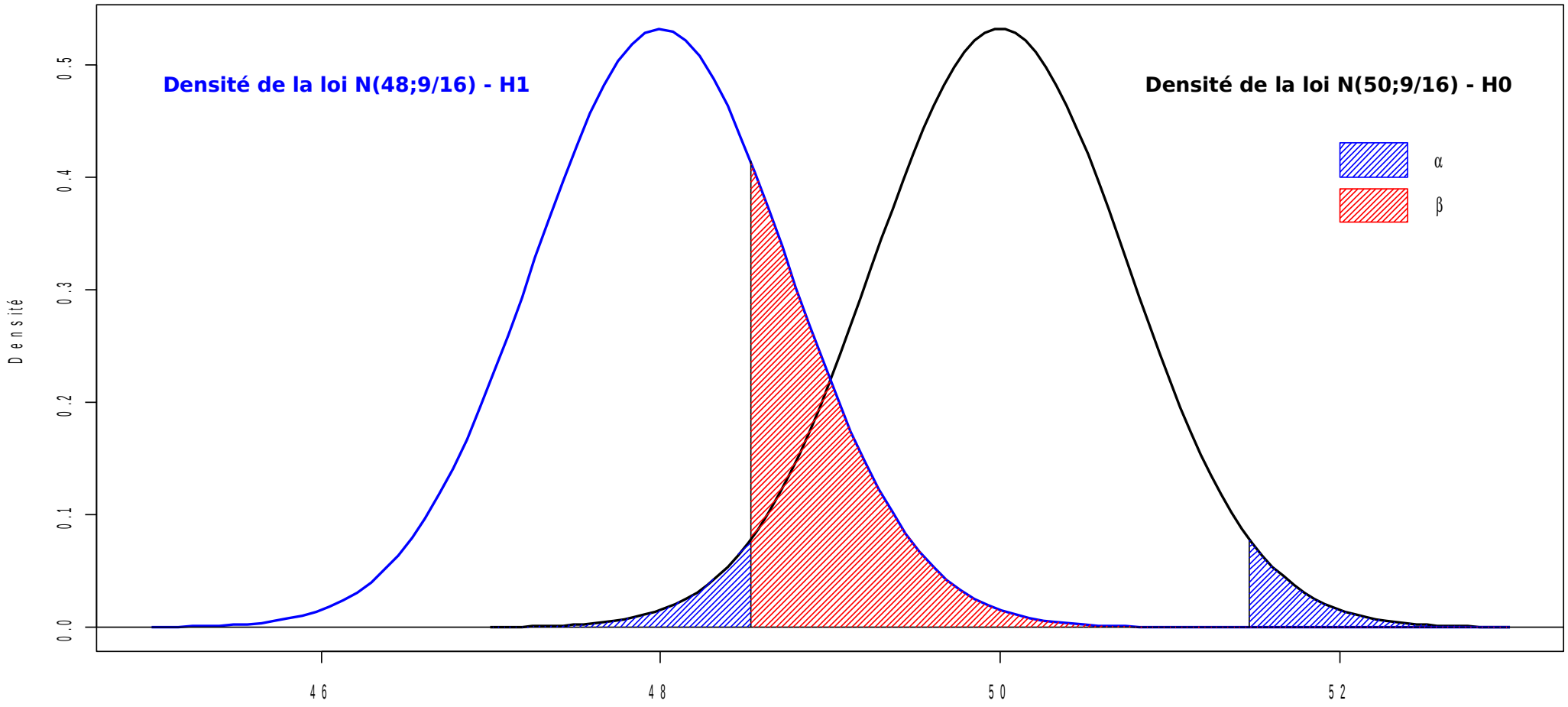
Représentation graphique de α et β

H0: μ=50
H1: μ=48

Rappel : pour α=5%, la région critique est]-∞ ; 48.53] ∪ [51.47 ; +∞ [

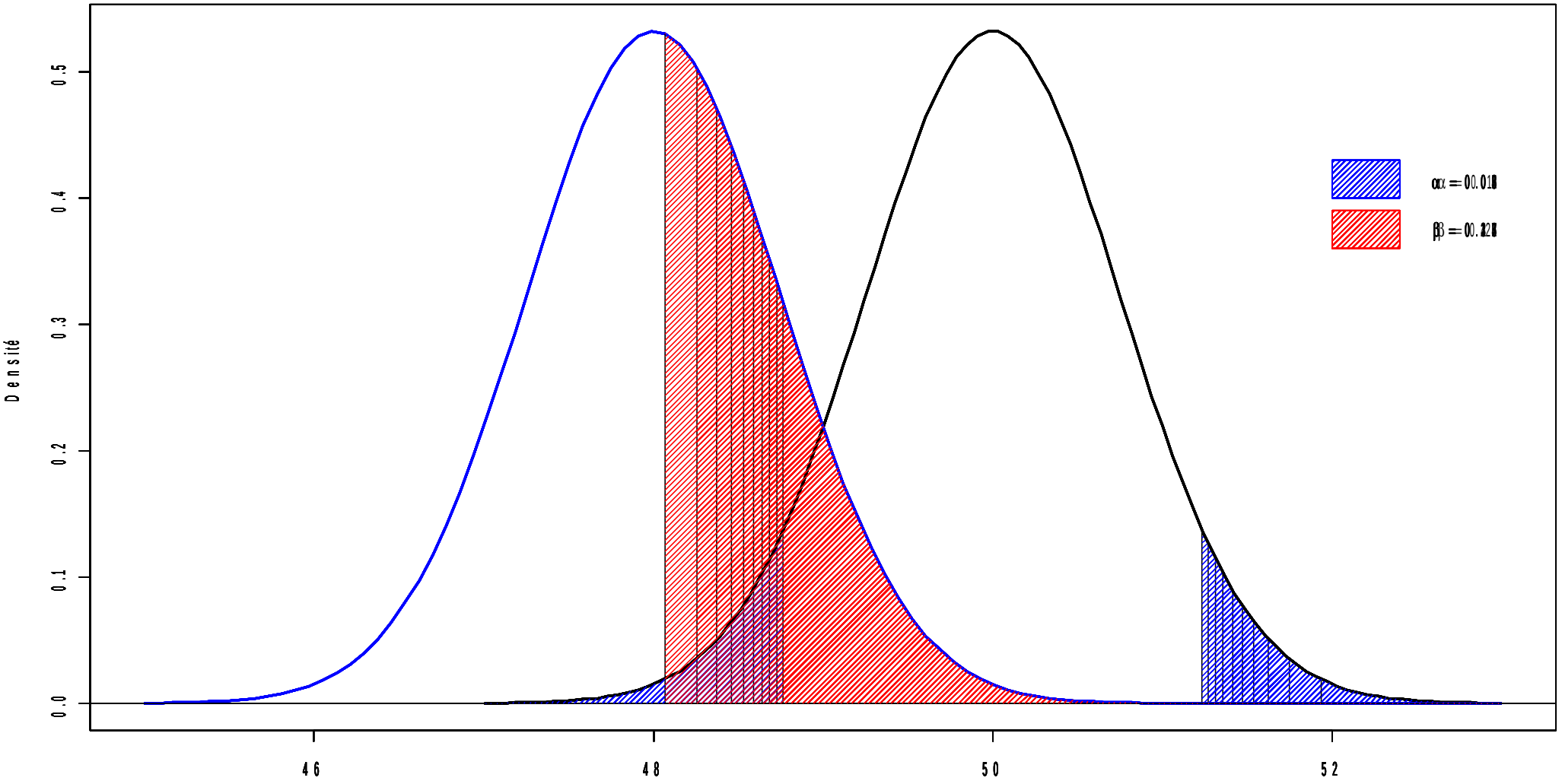
Dans ces conditions, β = 0.24

```
R> 1- pnorm(48.53,48,0.75)
```



Représentation graphique de α et β

Variations de α de 0.01 à 0.1

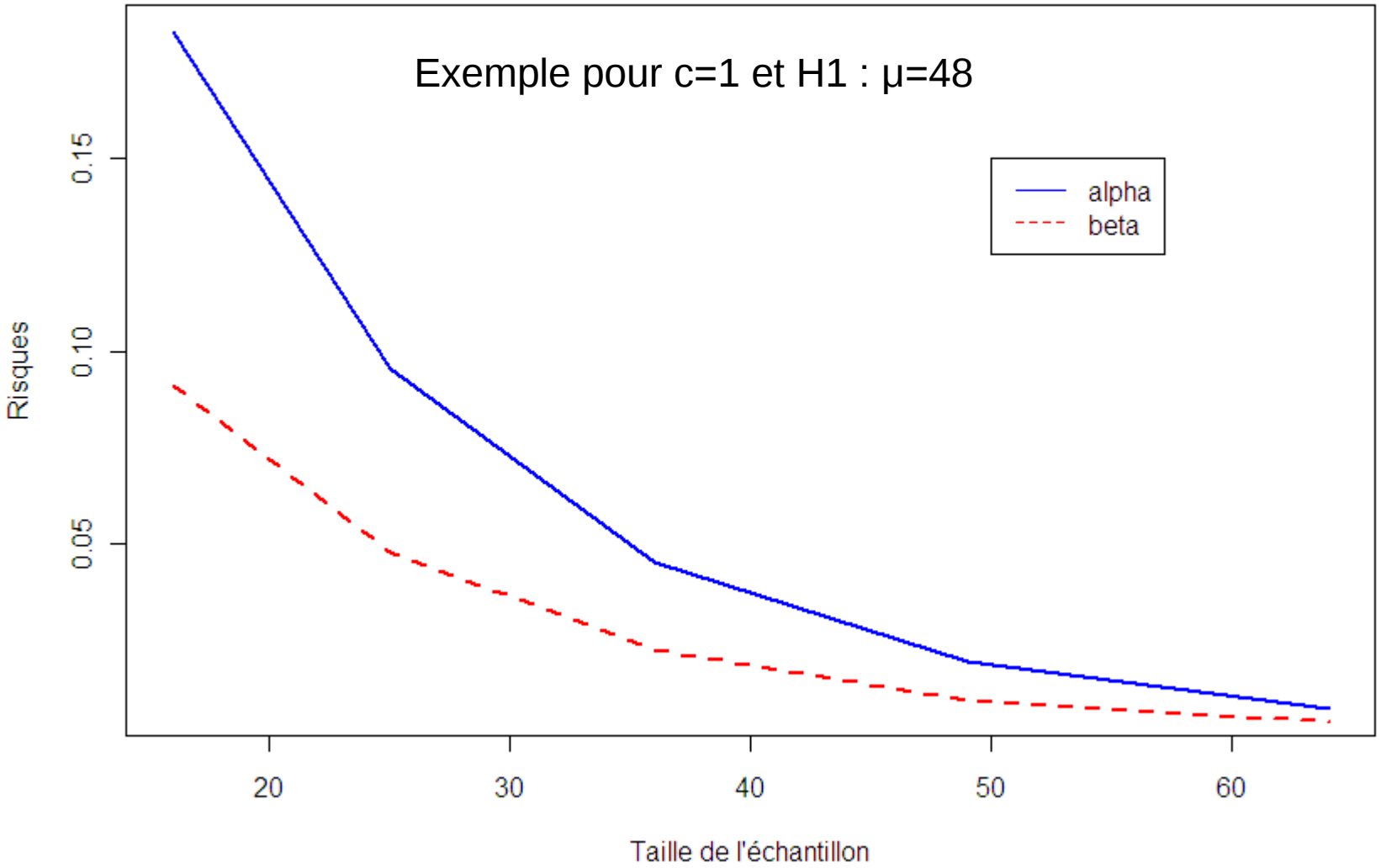


Test « significatif »

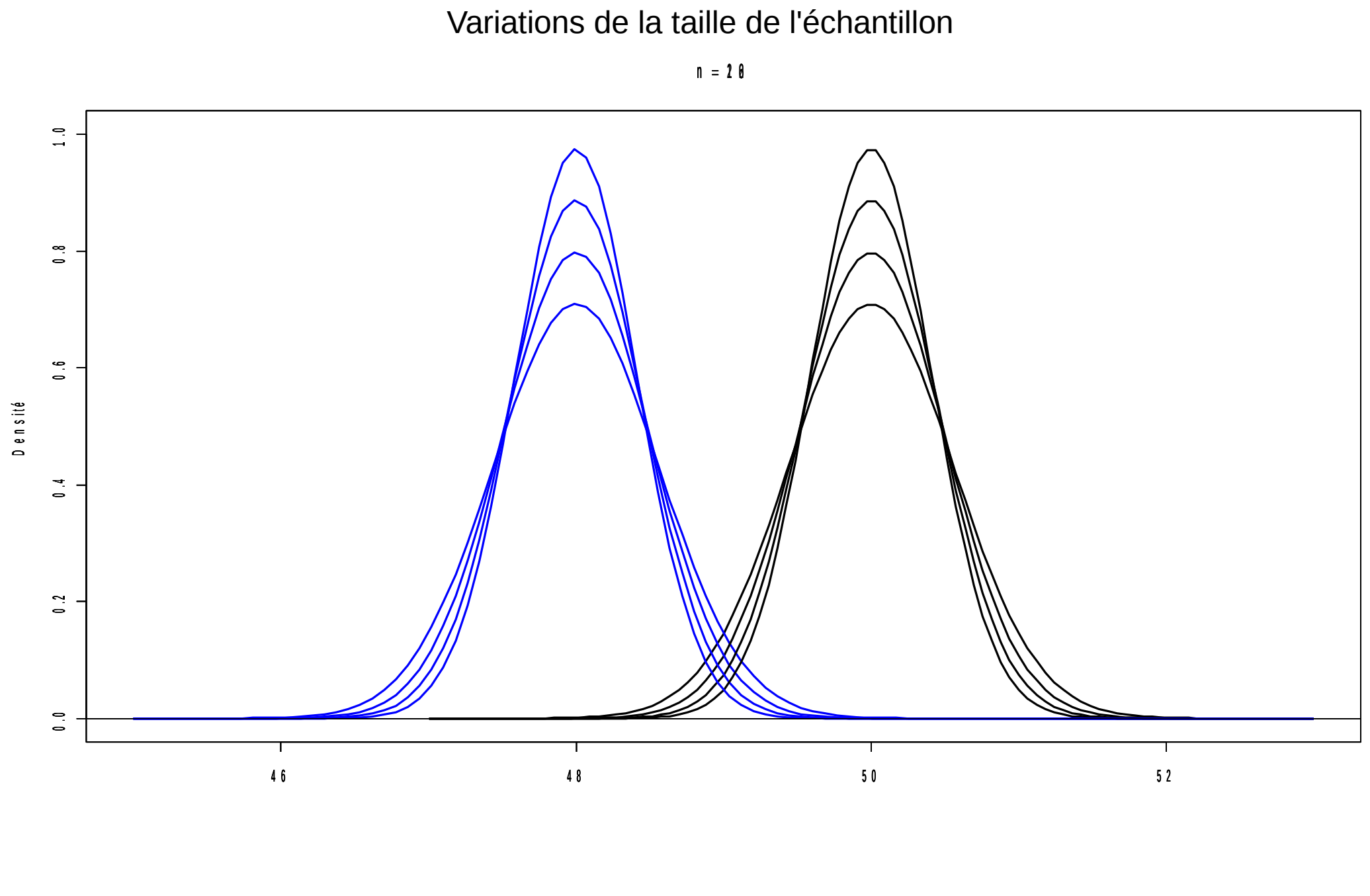
- Si le test conduit à rejeter H0, le risque de se tromper (α) est faible. La conclusion en faveur de H1 est solide. Le test est dit **significatif**.
- Si le test conduit à accepter H0,le risque de se tromper (β) peut être grand (selon l'hypothèse alternative). Cette conclusion est moins solide. Dans ce cas, le test est dit **non significatif**. D'où l'habitude d'affirmer « on ne peut pas rejeter H0 » plutôt que « on accepte H0 ».
- D'où la nécessaire réflexion du choix des hypothèses H0 et H1 ; H1 étant celle que l'on souhaite voir satisfaite avec un faible risque de se tromper.

Diminuer α et β

Le seul moyen de diminuer simultanément les risques α et β consiste à augmenter la taille de l'échantillon (ce qui implique une diminution de la variance de $\overline{X}$ et donc diminue le recouvrement des 2 courbes).



Représentation graphique de α et β



Calcul d'effectif : principe

La puissance diminue (= le risque β augmente) quand :

- Le risque α diminue et / ou
- La taille de l'échantillon diminue et / ou
- La taille de l'effet recherché diminue. En d'autres termes, l'hypothèse alternative H1 « se rapproche » de l'hypothèse nulle H0.

Calcul d'effectif : principe

Une fois fixés :

- Le risque α
- Le risque β ou la puissance du test (1-β)
- La taille de l'effet à mettre en évidence

seule la taille de l'échantillon reste comme inconnue.

On peut donc la définir **a priori** compte tenu des autres informations.

Calcul d'effectif : exemple

Extrait de l'aide en ligne de la fonction `power.t.test()` de R

```
Power calculations for one and two sample t tests

Description
Compute power of test, or determine parameters to obtain target power.

Usage
power.t.test(n = NULL, delta = NULL, sd = 1, sig.level = 0.05,
             power = NULL,
             type = c("two.sample", "one.sample", "paired"),
             alternative = c("two.sided", "one.sided"),
             strict = FALSE)

Arguments
n           Number of observations (per group)
delta       True difference in means
sd          Standard deviation
sig.level   Significance level (Type I error probability)
power       Power of test (1 minus Type II error probability)
type        Type of t test alternative One- or two-sided test
strict      Use strict interpretation in two-sided case

Details
Exactly one of the parameters n, delta, power, sd, and sig.level must be
passed as NULL, and that parameter is determined from the others.

...
```

Calcul d'effectif : exemple

R> power.t.test(n = 25, delta = 1, sd = 1, sig.level = 0.05, power = **NULL**)

```
     Two-sample t test power calculation

              n = 25
          delta = 1
             sd = 1
      sig.level = 0.05
          power = 0.9337076
    alternative = two.sided

 NOTE: n is number in *each* group
```

R> power.t.test(n = 30, delta = 1, sig.level = **NULL** ,power= 0.9)

```
              n = 30
          delta = 1
             sd = 1
      sig.level = 0.01286591
          power = 0.90
```

R> power.t.test(n = **NULL**, delta = 1, sd = 1, sig.level = 0.05, power = 0,9)

```
     Two-sample t test power calculation

              n = 22.02110
          delta = 1
             sd = 1
      sig.level = 0.05
          power = 0.9
    alternative = two.sided

 NOTE: n is number in *each* group
```

R> power.t.test(n = 30, delta = **NULL**, sig = 0.05 ,power= 0.9)

```
              n = 30
          delta = 0.8511743
             sd = 1
      sig.level = 0.05
          power = 0.9
```

Panorama de quelques tests statistiques

Problèmes à 1 échantillon

Type de test	Test paramétrique	Test non paramétrique
Conformité à 1 standard	Test de comparaison de moyenne (Student), d'écart-type, d'une proportion à une valeur de référence	
Adéquation à une loi		• Kolmogorov-Smirnov • χ^2 d'adéquation • Shapiro-Wilk

Association entre variables

Type de test	Test paramétrique	Test non paramétrique
2 variables quantitatives	Coefficient de corrélation de Pearson	
2 variables qualitatives		χ^2 d'indépendance

Panorama de quelques tests statistiques

Problèmes à K échantillons : comparaison de population

Type de test	Test paramétrique	Test non paramétrique
Comparaison de populations, les fonctions de répartition sont les mêmes dans les groupes		• **Kolmogorov-Smirnov** • Cramer – von Mises
Tests de comparaison de K échantillons indépendants (différenciation selon les caractéristiques de tendance centrale)	• **Test de comparaison de moyennes** (K=2) • **ANOVA** (analyse de variance) à 1 facteur	• **somme des rangs de Wilcoxon** (K=2) • **Mann - Whitney** (K=2) • Kruskal - Wallis • Test des médianes
Tests de comparaison de K échantillons indépendants (différenciation selon les caractéristiques de dispersion)	• **Fisher** (K=2) • Bartlett • Cochran • F-max de Hartley	• Ansari - Bradley • Siegel-Tukey • Test des différences extrêmes de Moses
Tests pour K échantillons appariés (mesures répétées ou blocs aléatoires complets)	• **Test de Student de comparaison de moyennes pour échantillons appariés** (K=2) • Test de comparaison de variances pour échantillons appariés (K=2) • ANOVA pour blocs aléatoires complets	• Test des signes (K=2) • **Rangs signés de Wilcoxon** (K=2) • Friedman • Test de McNemar (K=2, variables binaires) • Test Q de Cochran (variables binaires)
Tests multivariés pour K échantillons indépendants	• T^2 de Hotelling, comparaison de K=2 barycentres (vecteur des moyennes) • **MANOVA** (analyse de variance mutlivariée), comparaison de K barycentres : Lambda de Wilks, Trace de Pillai, Trace de Hotelling-Lawley, La plus grande valeur propre de Roy	

Données indépendantes ou appariées ?

- Données **indépendantes** : les observations sont indépendantes à l'intérieur de chaque échantillon et d'un échantillon à l'autre

Ex: résultats scolaires filles et garçons, dosage d'un produit chez 2 groupes de patients ayant reçu une molécule ou un placebo...

- Données **appariées** : les mêmes individus sont soumis à 2 mesures successives d'une même variable

Ex: notes de copies soumises à une double correction, dosage d'un produit avant et après un traitement chez les mêmes individus...

Test paramétrique ou non paramétrique ?

- Test **paramétrique** : les hypothèses nulle et alternative du test portent sur un paramètre statistique (moyenne ou variance par exemple). Ces tests nécessitent généralement des conditions de validité (distribution normale des données par exemple).

- Test **non paramétrique** : un test non paramétrique porte globalement sur la répartition des données sans hypothèse sur leur distribution.

Cas de 2 échantillons

« Comparaison de moyennes »

Type de test / Type de données	Test paramétrique	Test non paramétrique
Données indépendantes	Test de Student pour 2 échantillons	Test de Wilcoxon-Mann-Whitney *Rank-sum test*
Données appariées	Test de Student pour 1 échantillon (sur la différence)	Test de Wilcoxon *Signed-rank test*

Le test de Wilcoxon-Mann-Whitney (1)

Exemple : la concentration d'un produit est mesurée sur 2 échantillons indépendants de taille respective n_1=5 et n_2=6. Voici les mesures :

```
Ech 1 : 1.31 1.46 1.85 1.58 1.64
Ech 2 : 1.49 1.32 2.01 1.59 1.76 1.86
```

On souhaite savoir si les données sont significativement différentes dans les 2 groupes.

Procédure du test de W-M-W

1) Classer toutes les observations par ordre croissant
2) Affecter son rang à chaque observation
3) Calculer la somme des rangs d'un échantillon (en général celui de plus petite taille)

Mise en œuvre :

1) 1.31 1.32 1.46 1.49 1.58 1.59 1.64 1.76 1.85 1.86 2.01
2) 1 2 3 4 5 6 7 8 9 10 11
3) Somme des rangs en bleu : W_1 = **25**

L'hypothèse d'égalité des 2 distributions est rejetée si cette valeur s'éloigne « trop » d'une valeur « moyenne ».

Le test de Wilcoxon-Mann-Whitney (2)

Une procédure alternative (et équivalente) consiste à utiliser la statistique de test U liée à la précédente par la relation :

$$U = W_1 - n_1(n_1+1)/2$$

Elle correspond au nombre total de fois où un élément de l'échantillon 1 dépasse un élément de l'échantillon 2.

Illustration

1.31 1.32 1.46 1.49 1.58 1.59 1.64 1.76 1.85 1.86 2.01

- 1.85 est plus grand que 4 éléments de l'échantillon 2 (1.76 1.59 1.49 1.32)
- 1.64 est plus grand que 3 éléments de l'échantillon 2 (1.59 1.49 1.32)
- 1.58 est plus grand que 2 éléments de l'échantillon 2 (1.49 1.32)
- 1.46 est plus grand que 1 élément de l'échantillon 2 (1.32)

On obtient ainsi **U = 10** (ce qui est bien égal à 25-(5*6)/2).

C'est cette valeur que l'on retrouve dans la sortie de la fonction `wilcox.test()` du logiciel R.

La p-value obtenue ici (**`0.4286`**) indique qu'il n'y a pas de décalage (*shift*) entre les positions des 2 séries d'observations.

```
> x<-c(1.31,1.46,1.85,1.58,1.64)
> y<-c(1.49,1.32,2.01,1.59,1.76,1.86)
> wilcox.test(x,y)
        Wilcoxon rank sum test
data:  x and y
W = 10, p-value = 0.4286
alternative hypothesis: true location
shift is not equal to 0
```

Le test de Student

On considère le même problème que précédemment et on applique un test de Student pour comparer la moyenne des 2 échantillons même si les conditions d'application sont plus que discutables.

Calculs

	1.31	1.49
	1.46	1.32
	1.85	2.01
	1.58	1.59
	1.64	1.76
		1.86
Moyenne	1.658	1.672
Variance	0.041	0.064
Var. Commune	0.054	

t = -0.738

Formules

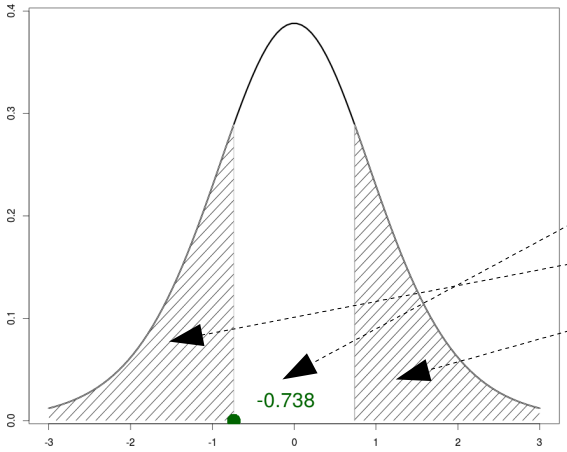
Sous H0, hypothèse d'égalité des moyennes, on a :

$$\frac{\bar{x}-\bar{y}}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim Student\left(n_1+n_2-2\right)$$

Avec s^2 la variance commune aux 2 échantillons

$$s^2=\frac{(n_1-1)V_1+(n_2-1)V_2}{n_1+n_2-2}$$

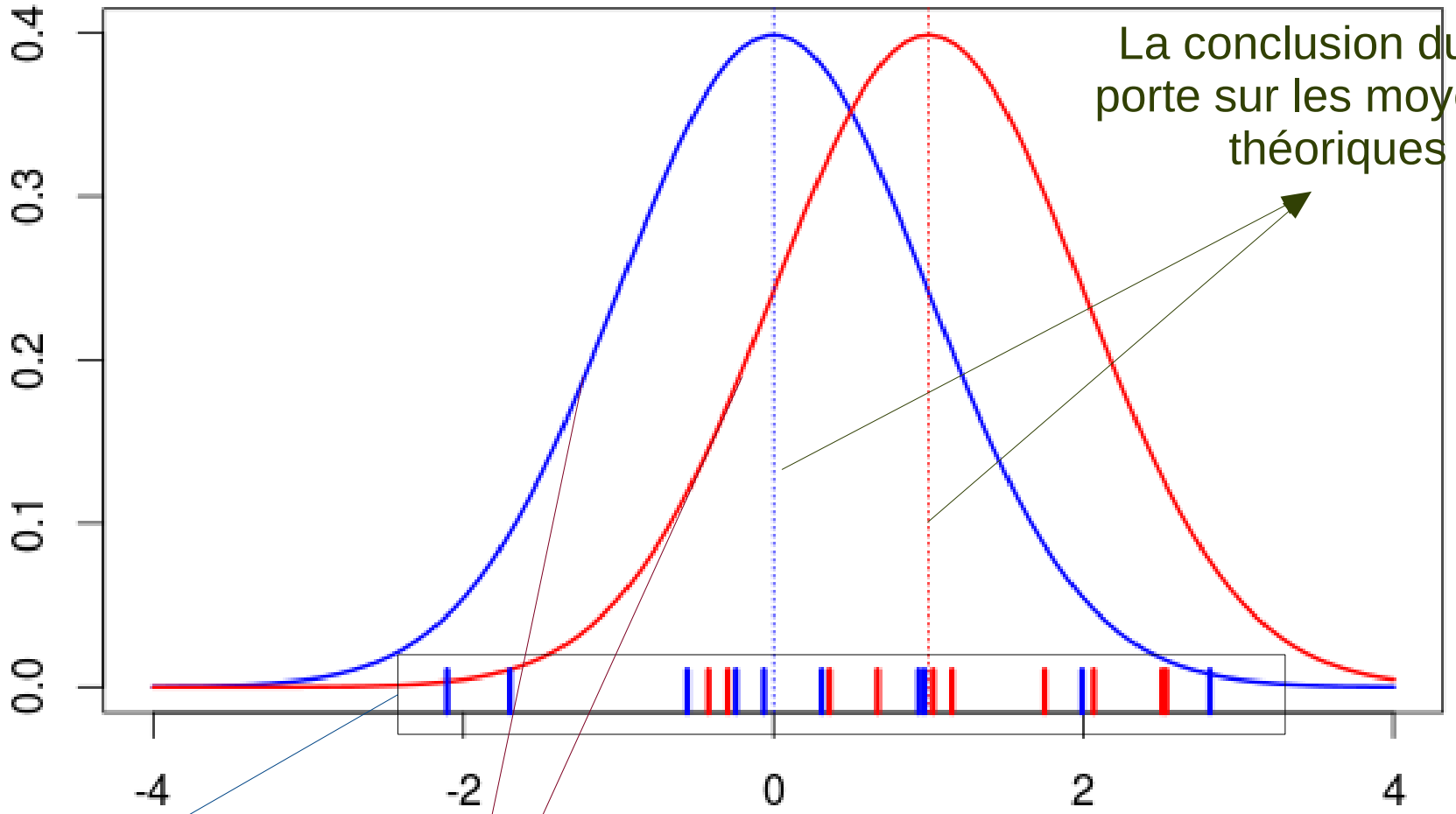
Densité de la loi de Student à 9 ddl




```
> t.test(x,y,var.equal=T)
        Two Sample t-test
data:  x and y
t = -0.7381, df = 9, p-value = 0.4792
alternative hypothesis: true difference
in means is not equal to 0
95 percent confidence interval:
 -0.4213783  0.2140450
sample estimates:
mean of x mean of y
 1.568000  1.671667
```


D

Le test de Student



La conclusion du test porte sur les moyennes théoriques

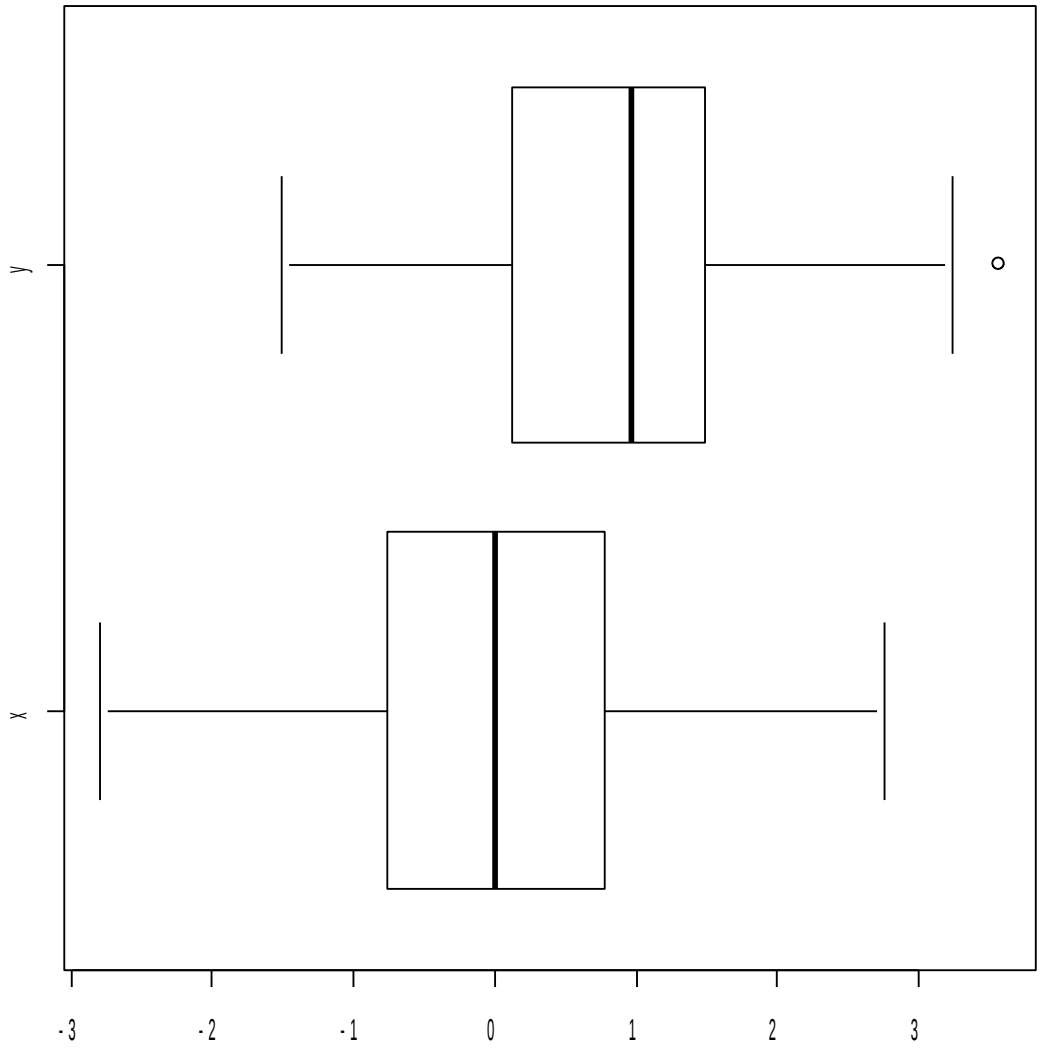
Observations

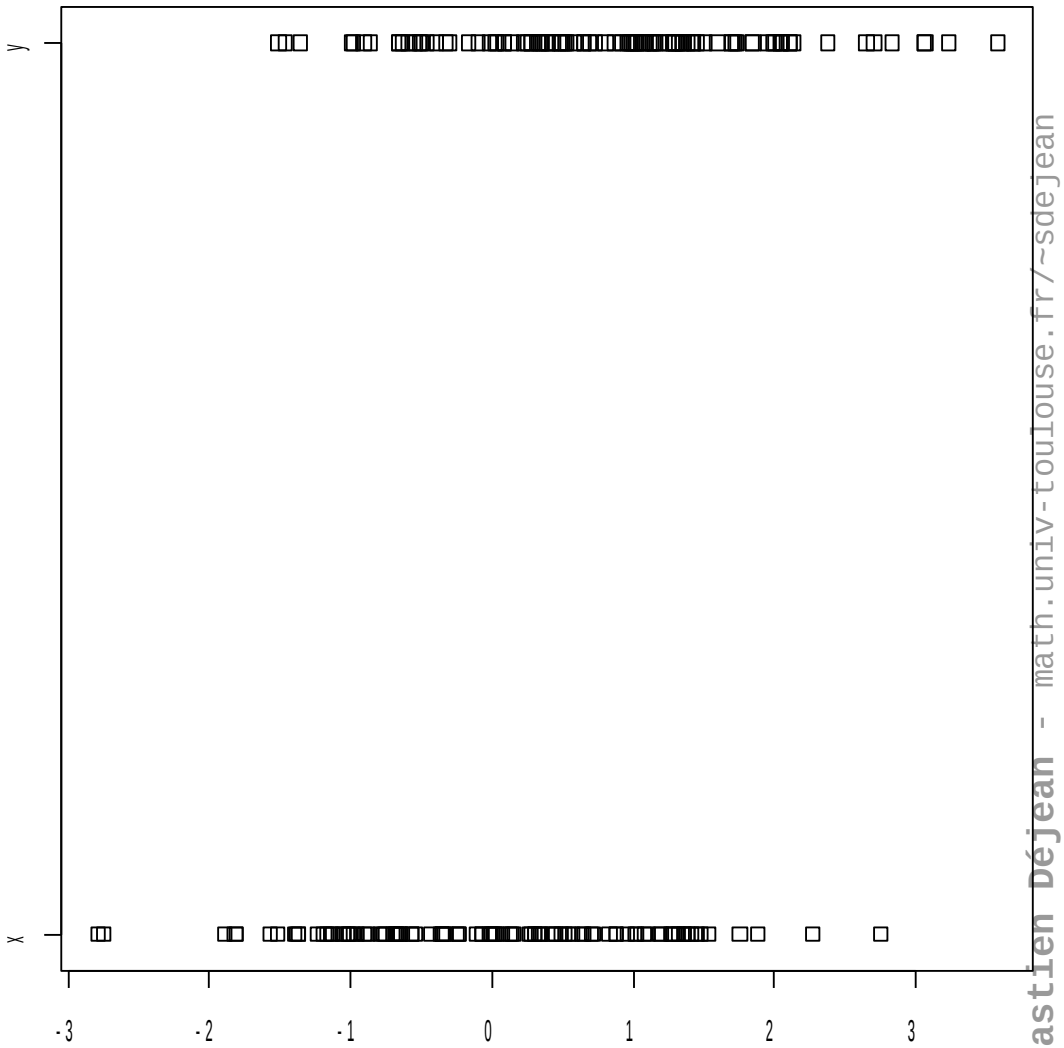
Hypothèses nécessaires à la réalisation du test

D

Mise en œuvre de quelques tests

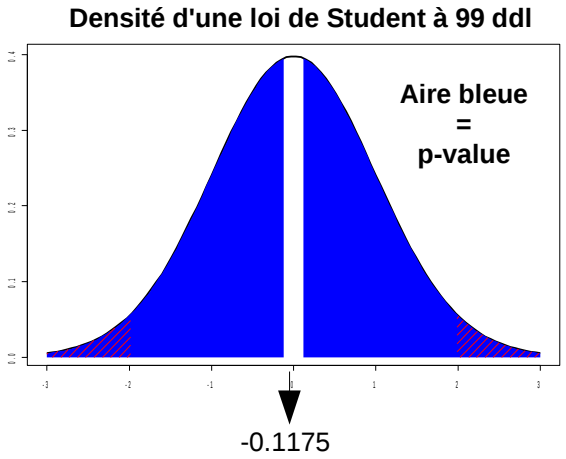
Données simulées : génération aléatoire selon une loi normale de 2 échantillons de longueur 100 : x ~ N(0,1) et y ~ N(1,1)





Mise en œuvre de quelques tests

Test de Student pour un échantillon

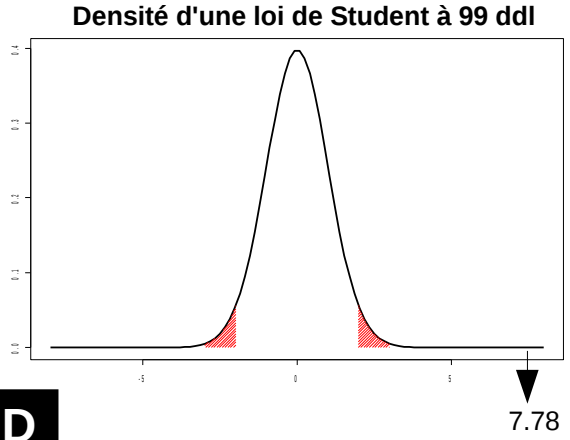


On ne peut pas rejeter H0, la moyenne est probablement nulle.

```
        One Sample t-test

data:  x
t = -0.1175, df = 99, p-value = 0.9067
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.2239679  0.1989233
sample estimates:
  mean of x
-0.01252230
```

Rejet de H0 avec une très faible probabilité de se tromper.




```
        One Sample t-test

data:  y
t = 7.78, df = 99, p-value = 7.082e-12
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.6291375 1.0599157
sample estimates:
mean of x
0.8445266
```

Mise en œuvre de quelques tests

Test de Fisher d'égalité des variances

```
        F test to compare two variances

data:  x and y
F = 0.9637, num df = 99, denom df = 99, p-value = 0.8545
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.6484291 1.4323091
sample estimates:
ratio of variances
         0.9637173
```

On ne peut pas rejeter H0, les 2 variances sont très probablement égales.

Mise en œuvre de quelques tests

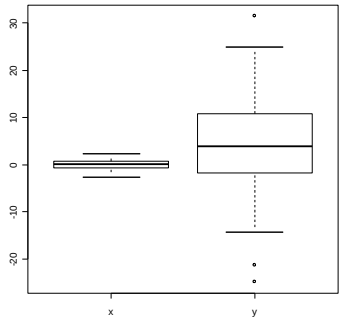
Test de Student pour 2 échantillons

Two Sample t-test (*variances supposées égales*)

```
data:  x and y
t = -5.6342, df = 198, p-value = 5.982e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.1570238 -0.5570741
sample estimates:
  mean of x   mean of y
-0.01252230  0.84452662
```

On rejette H0, les 2 moyennes sont très probablement différentes.

Pour effectuer ce test, on suppose les 2 variances égales. Cela peut être contrôlé par un test de Fisher d'égalité des variances. Dans le cas ci-contre, la comparaison des moyennes n'a pas vraiment de sens.



Welch Two Sample t-test

adaptation du test de Student sans l'hypothèse de variances égales

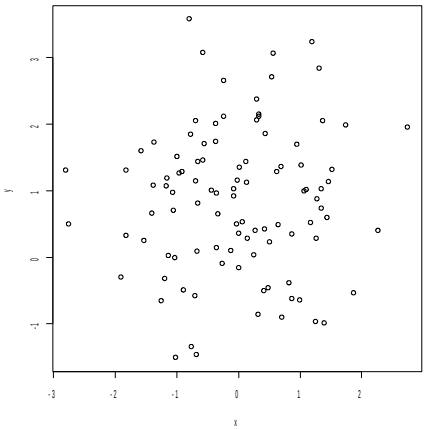
```
data:  x and y
t = -5.6342, df = 197.932, p-value = 5.985e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.1570244 -0.5570734
sample estimates:
  mean of x   mean of y
-0.01252230  0.84452662
```

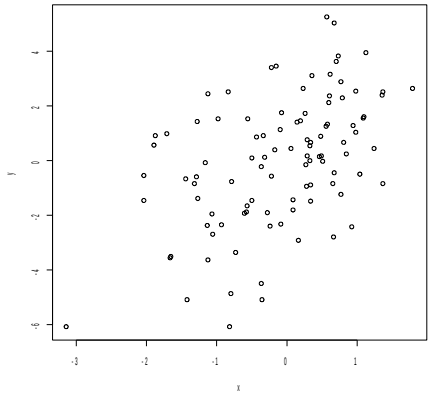
D

Mise en œuvre de quelques tests

Test sur le coefficient de corrélation

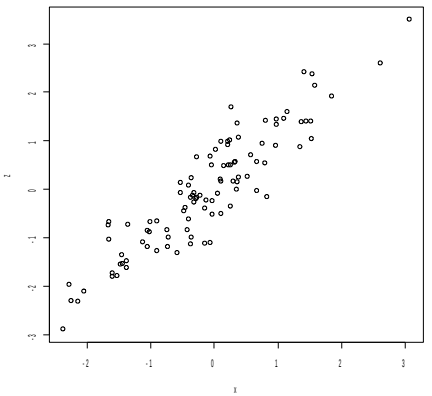
```
        Pearson's product-moment correlation
data:  x and y
t = 0.5464, df = 98, p-value = 0.586
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.1428544  0.2488346
sample estimates: cor = 0.05511005
```





```
        Pearson's product-moment correlation
data:  x and z1
t = 5.5115, df = 98, p-value = 2.88e-07
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3206572 0.6233025
sample estimates: cor = 0.486438
```

```
        Pearson's product-moment correlation
data:  x and z2
t = 22.3231, df = 98, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8748002 0.9415099
sample estimates: cor = 0.914144
```



Mise en œuvre de quelques tests

Test de normalité Kolmogorov-Smirnov

```
> ks.test(x,y) # x et Y sont-ils des échantillons d'une même distribution ?
        Two-sample Kolmogorov-Smirnov test
data:  x and y
D = 0.33, p-value = 3.729e-05
alternative hypothesis: two-sided
```

Probablement pas, avec une faible chance de se tromper

```
> ks.test(x,"pnorm") # x est-il un échantillon d'une loi normale N(0,1) ?
        One-sample Kolmogorov-Smirnov test
data:  x
D = 0.0718, p-value = 0.6803
alternative hypothesis: two-sided
```

Les données ne permettent pas de dire le contraire.

```
> ks.test(y,"pnorm") # y est-il un échantillon d'une loi normale N(0,1) ?
        One-sample Kolmogorov-Smirnov test
data:  y
D = 0.3408, p-value = 1.641e-10
alternative hypothesis: two-sided
```

Probablement pas, avec une très faible chance de se tromper.

```
> ks.test(y,"pnorm",1) y est-il un échantillon d'une loi normale N(1,1) ?
        One-sample Kolmogorov-Smirnov test
data:  y
D = 0.0923, p-value = 0.3614
alternative hypothesis: two-sided
```

Les données ne permettent pas de dire le contraire.

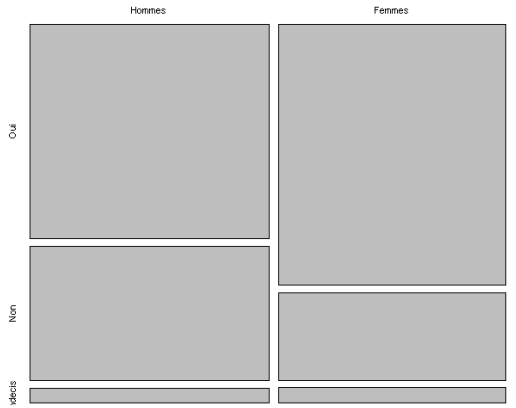
Quelques tests classiques

Test du χ^2 d'indépendance

- **Données** : effectifs recueillies dans une table de contingence (tableau croisé pour 2 variables qualitatives)
- **Question** : les 2 variables qualitatives sont-elles indépendantes ?

- **Exemple** : 1250 personnes ont répondu à la question « Êtes-vous satisfaits des programmes TV ? ». On souhaite savoir si la satisfaction dépend du sexe.

	OUI	NON	Indécis	**Somme**
Hommes	378	237	26	**641**
Femmes	438	146	25	**609**
Somme	**816**	**383**	**51**	**1250**



Hypothèse H0 : Satisfaction et sexe sont indépendants

Effectifs théoriques sous l'hypothèse d'indépendance
*(effectif d'une case = effectif de la ligne * effectif de la colonne / effectif total)*

	OUI	NON	Indécis
Hommes	418	196	26
Femmes	398	187	25

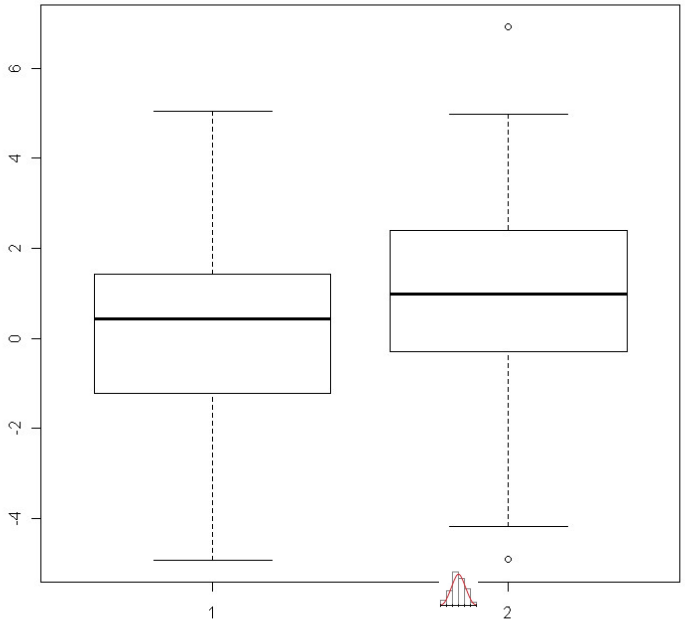
Statistique de test : χ^2_{obs} = Σ (Obs-Théo)^2 / Théo

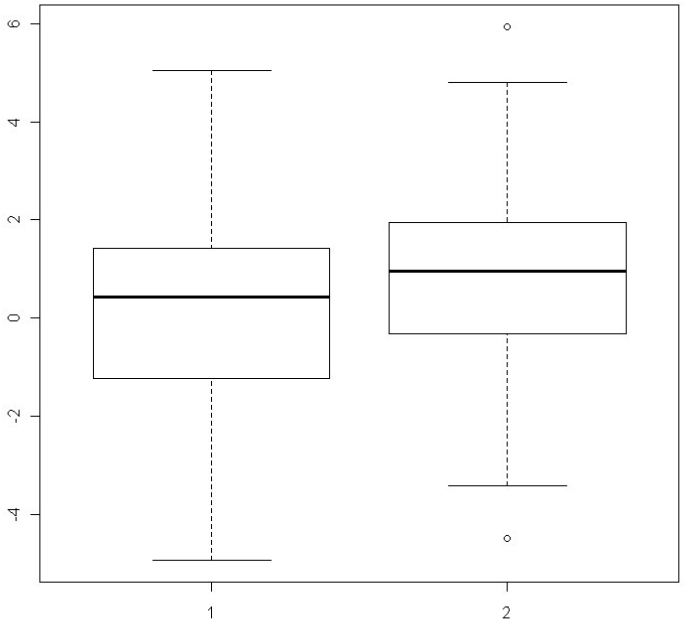
Reflète l'écart entre les données observées et les effectifs théoriques en cas d'indépendance

```
Pearson's Chi-squared test
data:  data.chisq
X-squared = 25.2501, df = 2, p-value = 3.289e-06
```

Les 2 caractères ne semblent pas indépendants.

P-value et *





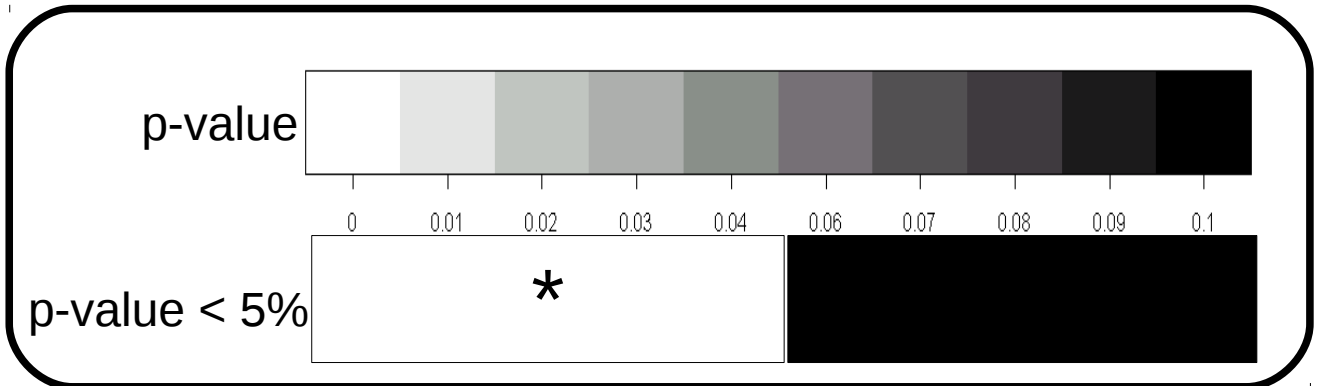
`Two Sample t-test : p-value =` **`0.08284`**

`Two Sample t-test : p-value =` **`0.03556`**

P-value > 5%

P-value < 10% *

P-value < 5% *



D

Une petite liste de tests « courants »

Extrait de l'aide en ligne du logiciel R (sans package additionnel)

stats::ansari.test	Ansari-Bradley Test
stats::bartlett.test	Bartlett Test of Homogeneity of Variances
stats::binom.test	Exact Binomial Test
stats::Box.test	Box-Pierce and Ljung-Box Tests
stats::chisq.test	Pearson's Chi-squared Test for Count Data
stats::cor.test	Test for Association/Correlation Between Paired Samples
stats::fisher.test	Fisher's Exact Test for Count Data
stats::fligner.test	Fligner-Killeen Test of Homogeneity of Variances
stats::friedman.test	Friedman Rank Sum Test
stats::kruskal.test	Kruskal-Wallis Rank Sum Test
stats::ks.test	Kolmogorov-Smirnov Tests
stats::mantelhaen.test	Cochran-Mantel-Haenszel Chi-Squared Test for Count Data
stats::mauchly.test	Mauchly's Test of Sphericity
stats::mcnemar.test	McNemar's Chi-squared Test for Count Data
stats::mood.test	Mood Two-Sample Test of Scale
stats::oneway.test	Test for Equal Means in a One-Way Layout
stats::pairwise.prop.test	Pairwise comparisons for proportions
stats::pairwise.t.test	Pairwise t tests
stats::pairwise.wilcox.test	Pairwise Wilcoxon rank sum tests
stats::poisson.test	Exact Poisson tests
stats::power.anova.test	Power calculations for balanced one-way analysis of variance tests
stats::power.prop.test	Power calculations two sample test for proportions
stats::power.t.test	Power calculations for one and two sample t tests
stats::PP.test	Phillips-Perron Test for Unit Roots
stats::prop.test	Test of Equal or Given Proportions
stats::prop.trend.test	Test for trend in proportions
stats::quade.test	Quade Test
stats::shapiro.test	Shapiro-Wilk Normality Test
stats::t.test	Student's t-Test
stats::var.test	F Test to Compare Two Variances
stats::wilcox.test	Wilcoxon Rank Sum and Signed Rank Tests

Problème de la multiplicité

La **roulette russe** est un jeu consistant à mettre une ou plusieurs cartouches (suivant la probabilité souhaitée) dans le barillet d'un revolver, à tourner ce dernier de manière aléatoire puis à pointer le revolver sur sa tempe avant d'actionner la détente. Si une cartouche se trouve alors dans la chambre placée dans l'axe du canon elle sera percutée et le « joueur » mourra ou sera blessé. (Wikipedia)

Supposons que le barillet contienne **100** emplacements de cartouche et que le « joueur » mette **5** cartouches aléatoirement. Sa probabilité d'être blessé (soyons optimiste) est de 5%.

Supposons maintenant que l'individu « joue » plusieurs fois de suite. Intuitivement, on sent bien qu'il va finir par « se faire mal » et que la probabilité de « perdre » au moins une fois augmente à chaque tentative.

Nous allons traduire mathématiquement cette petite histoire.

Problème de la multiplicité

Calculons la probabilité qu'à le « joueur » de rester bien portant après n tentatives

- Pour n=1, $P[bp_{(1)}]=1-0.05=\mathbf{0.95}$
- Pour n=2, il faut que le « joueur » soit bien portant après le 1er essai **ET** après le second

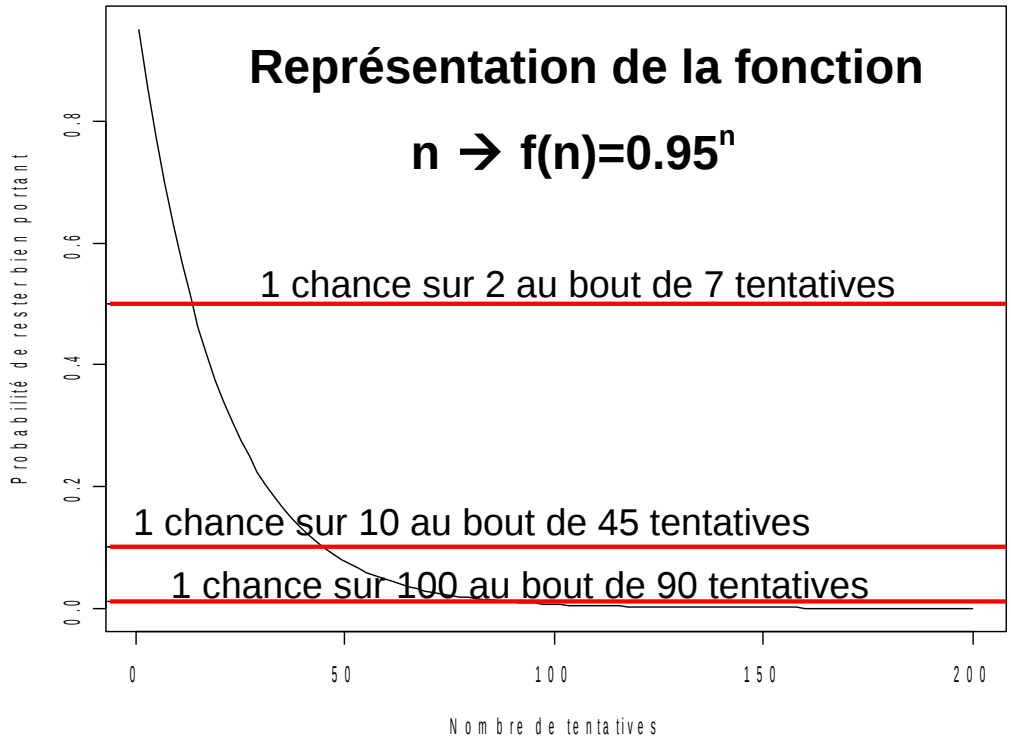
$$P[bp_{(2)}]=(1-0.05)*(1-0.05)=0.95^2=\mathbf{0.9025}$$

- Et ainsi de suite…

Règle générale : $P[bp_{(n)}]=(1-0.05)^n$

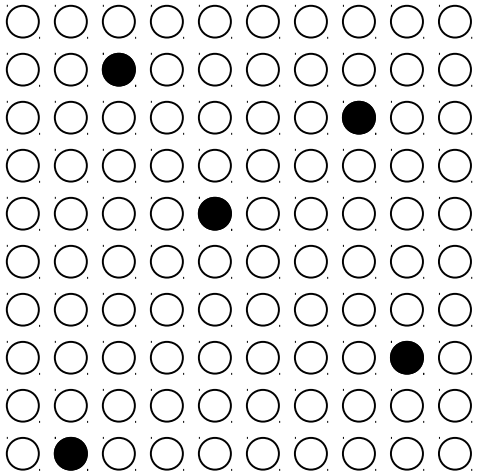
Quelques valeurs

n	**1**	**2**	**5**	**10**	**100**
P[bp(n)]	0.95	0.902	0.774	0.599	0.006



Problème de la multiplicité

Une autre façon de considérer le problème consiste à envisager plusieurs « joueurs » jouant simultanément. Si le risque est toujours fixé à 5% (avec 5 cartouches sur 100 emplacements), et si 100 joueurs actionne la détente, on peut s'attendre à ce que 5 joueurs perdent. Ils seront 50 pour 1000 joueurs…



Cette façon de voir le problème de multiplicité est plus proche de la problématique de la détection de gènes différentiellement exprimés. En effet, un test statistique effectué pour l'ensemble des tests d'une biopuce (plusieurs milliers) au risque de 5% implique nécessairement la détection de gènes non pertinents.

Illustration : dans le cas de l'analyse des biopuces, la conséquence de l'erreur est moins radicale que pour la roulette russe.

Si les tests sont effectués pour détecter les gènes différentiellement exprimés (DE) entre 2 conditions alors le risque consiste à déclarer un gène DE alors qu'il ne l'est pas : on parle de gène faux-positif (FP). Au contraire, un gène « réellement » DE non détecté par le test est dit faux-négatif (FN).

Les conséquences de ces 2 types d'erreur sont de natures diverses. Dans le premier cas, on va engager des moyens (humains et financiers) pour valider l'hypothèse que le gène FP est effectivement intéressant alors qu'il ne l'est pas. Dans le second cas (FN), on passe peut-être à côté de la découverte du siècle en négligeant un gène important !!!

Exemple de tests multiples

- 1 comparaison : α = P[Rejeter H0 // H0 vraie] = 5%

  ➔ Probabilité de ne pas commettre d'erreur : 1-0.05 = 0.95

- 3 comparaisons : $\begin{cases} \alpha_1 = P[\text{Rejeter } H0_1 \,//\, H0_1 \text{ vraie}] = 5\% \\ \alpha_2 = P[\text{Rejeter } H0_2 \,//\, H0_2 \text{ vraie}] = 5\% \\ \alpha_3 = P[\text{Rejeter } H0_3 \,//\, H0_3 \text{ vraie}] = 5\% \end{cases}$

➔ Probabilité de ne pas commettre d'erreur = produit des probabilités de ne pas commettre d'erreur à chacune des 3 comparaisons = (1-0.05) * (1-0.05) * (1-0.05) = 0.86

Le risque (global) de commettre au moins une erreur est :

1-0.86 = 0.14

Il faut donc diminuer le risque associé à chaque comparaison pour contrôler le risque global.

Exemple de tests multiples

Méthode de Bonferroni

La méthode de Bonferroni consiste à corriger le risque associé à chaque comparaison (α_i) en le divisant par le nombre de comparaisons à effectuer pour contrôler le risque global (α_g)

En prenant $\alpha_1=\alpha_2=\alpha_3=\alpha_g/3=0.05/3$, on a la probabilité de ne pas commettre d'erreur :

$$(1-0.05/3) * (1-0.05/3) * (1-0.05/3) = 0.9508$$

Ce qui correspond à un risque global inférieur à 5%.

Dans le cas des biopuces comportant plusieurs centaines voire plusieurs milliers de gènes, cette correction peut devenir trop conservative et conduire à la détection d'aucun gène. D'autres méthodes plus sophistiquées existent : Sidak, Holm, Westfall et Young, Hochberg…

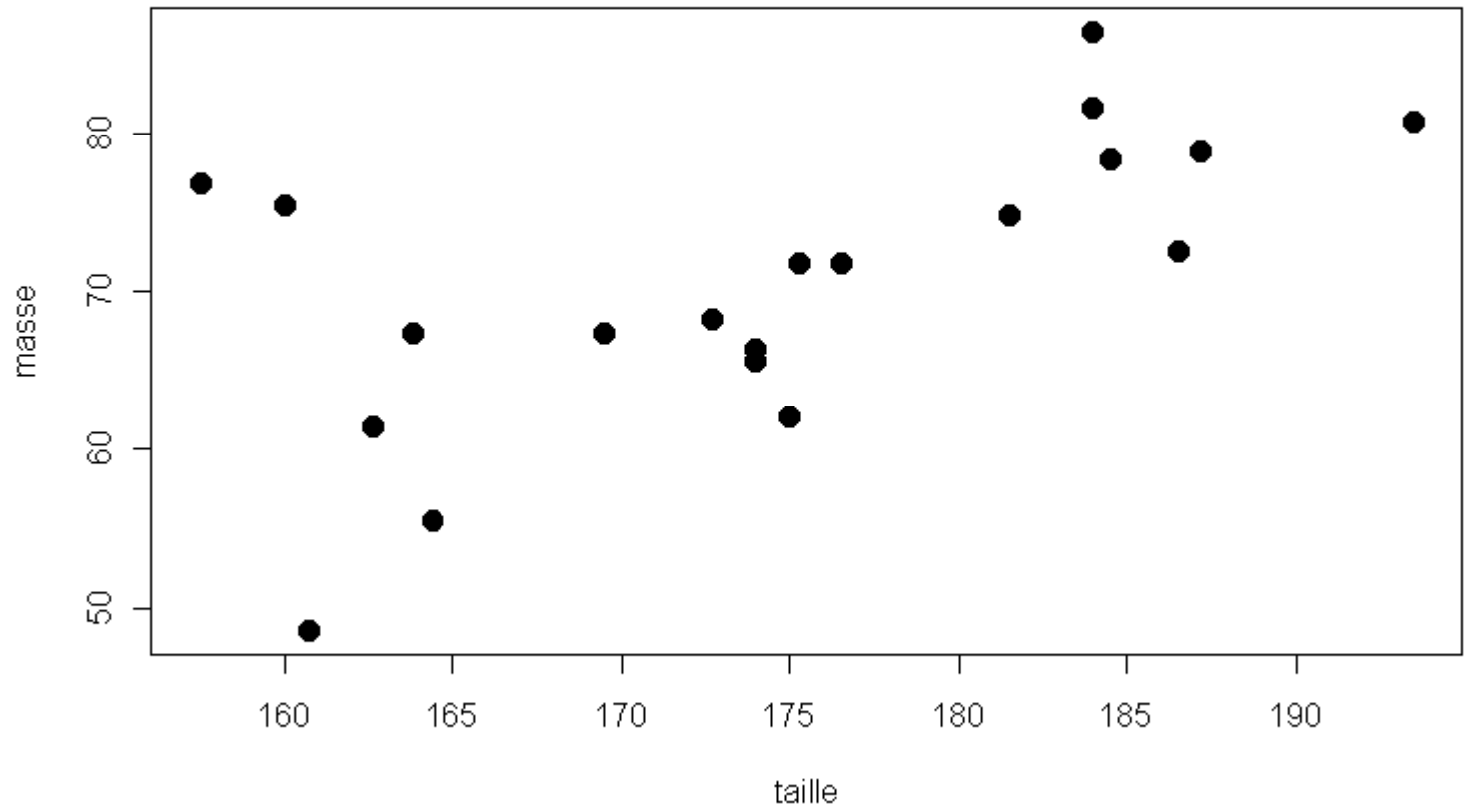
Modélisation

1 ou plusieurs variables **à expliquer**
par
1 ou plusieurs variables
explicatives

- Régression linéaire simple :
  - **1** variable à expliquer **quantitative** et **1** variable explicative **quantitative**
- ANOVA 1 facteur
  - **1** variable à expliquer **quantitative**, **1** variable explicative **qualitative** (facteur)

Régression linéaire simple

Taille (cm) :	174.0	175.3	193.5	186.5	187.2	181.5	184.0	184.5	175.0	184.0	169.5	160.0	172.7	162.6	157.5	176.5	164.4	160.7	174.0	163.8
Masse (kg) :		65.6	71.8	80.7	72.6	78.8	74.8	86.4	78.4	62.0	81.6	67.3	75.5	68.2	61.4	76.8	71.8	55.5	48.6	66.4
67.3																				



Peut-on « modéliser » « correctement » par une droite la masse des individus en fonction de la taille ?

Régression linéaire simple

Équation d'une droite : $Y = aX + b + \varepsilon$

Comment déterminer a et b ?

Par exemple, critère des moindres carrés : trouver a et b qui minimisent

$$\Sigma_i (y_i - ax_i - b)^2 = \Sigma_i \varepsilon_i^2$$

Sur l'exemple : a = 0.5445 ; b = -24.37

Un individu qui mesure 180cm pèse, selon le modèle, 180*0.5445-24.37 = 73.6kg

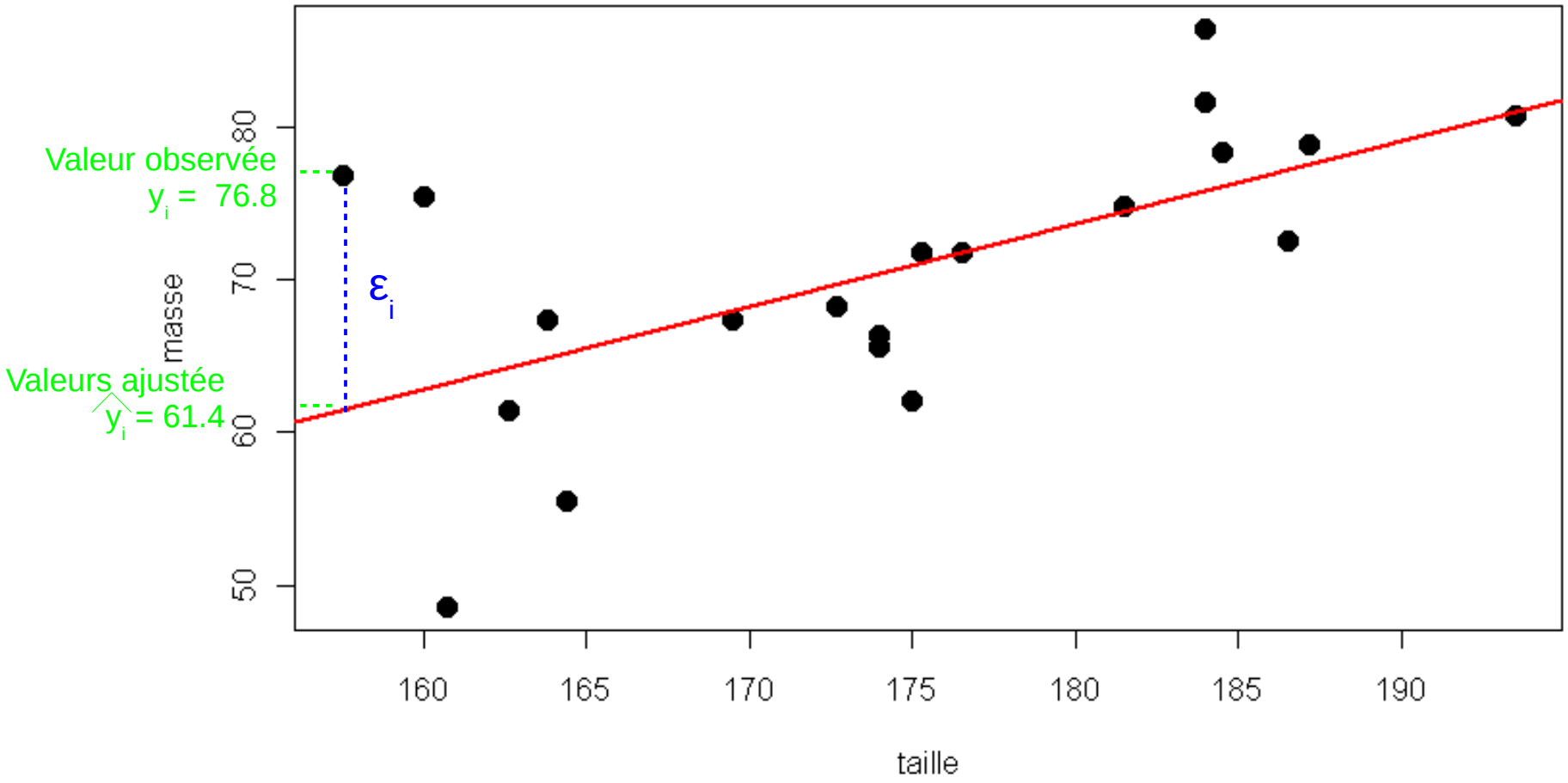
Régression linéaire simple

Équation de la droite :

$Y = 0.5445\ X - 24.37$

$R^2 = 0{,}3782 = var(y_i) / var(\hat{y}_i)$

observés ajustés



ANOVA 1 facteur

Notations

Modalité	Répétitions				Moyenne
	1	2	…	J	
1	Y_{11}	Y_{12}	…	Y_{1J}	$\overline{Y}_{1.}$
2	Y_{21}	Y_{22}	…	Y_{2J}	$\overline{Y}_{2.}$
…	…	…		…	
I	Y_{I1}	Y_{I2}	…	Y_{IJ}	$\overline{Y}_{I.}$
					$\overline{Y}_{..}$

- Y_{ij} : $j^{ème}$ observation pour la $i^{ème}$ modalité du facteur
- $\overline{Y}_{i.}$: moyenne des observations de la $i^{ème}$ modalité
- $\overline{Y}_{..}$: moyenne générale
- I modalités, J observations par modalité (plan équilibré)
- N=IJ observations

ANOVA 1 facteur

Modèle et sommes de carrés

Modèle d'ANOVA : $$Y_{ij} = \mu_i + \varepsilon_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

μ : effet moyen général

α_i : effet différentiel de la i^{ème} modalité du facteur

Test de l'effet du facteur : $$H0 : \forall i = 1,...,I \quad \alpha_i = 0$$

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\left(y_{ij} - \bar{y}_{..}\right)^2 = J\sum_{i=1}^{I}\left(\bar{y}_{i.} - \bar{y}_{..}\right)^2 + \sum_{i=1}^{I}\sum_{j=1}^{J}\left(y_{ij} - \bar{y}_{i.}\right)^2$$

SC_{totale} $SC_{facteur}$ $SC_{résidu}$

ANOVA 1 facteur

Table d'ANOVA

Source de variation	Degrés de liberté	Somme de carrés	Carrés moyens	Statistique de test F	P value
Variation « facteur »	I-1	$SC_{facteur}$	$CM_f = SC_f/(I-1)$	$F = CM_f/CM_r$	...
Variation « résidu »	N-I	$SC_{résidu}$	$CM_r = SC_r/(N-I)$		
Variation totale	N-1	SC_{totale}			

Règle de décision :

rejet de H0 $\Leftrightarrow$ $F > \mathbf{F}_{I-1,N-I,\alpha}$

Plus la variabilité due au facteur est importante par rapport à celle due au résidu, plus le test est susceptible de conclure à un effet du facteur.

Degré de liberté

Dans la table d'ANOVA présentée précédemment, on retrouve une colonne « degré de liberté ». Cette information est essentielle dans tous les aspects de la statistique liés à la modélisation et aux tests. Elle traduit la quantité d'information dont on dispose pour estimer une quantité inconnue. Dans le calcul des degrés de liberté, on se retrouve systématiquement avec une valeur **–1**.

→ Imaginons **3** personnes (Pierre, Paul, Jacques) face à **3** vêtements **Rouge**, **Vert** ou **Bleu**. Chacune des personnes doit choisir à son tour un vêtement.

- Pierre commence et a le choix entre **3** vêtements **R**, **V**, **B**. Il choisit **R**.
- Paul vient ensuite. Il ne reste que **2** choix possibles : **V** et **B**. Il choisit **B**.
- Jacques arrive et **n'a plus le choix**, il prend **V** !

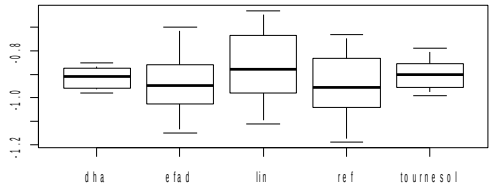
Ainsi, dans une situation où **3** unités sont initialement présentes, on constate que seulement **2** choix sont possibles ; le troisième étant nécessairement imposé.

→ Autre exemple : au moment de remplir le formulaire d'évaluation de la formation, vous aurez probablement la possibilité de rester anonyme. Si tel est votre souhait, assurez-vous qu'une autre personne le souhaite également. Sinon avec l'identité des **n-1** autres personnes, votre anonymat ne sera pas trop difficile à lever.

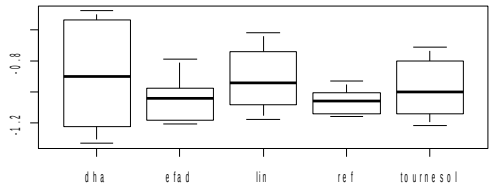
Mise en œuvre de l'ANOVA

Exemple : mesure de 4 variables quantitatives selon les 5 modalités d'un facteur

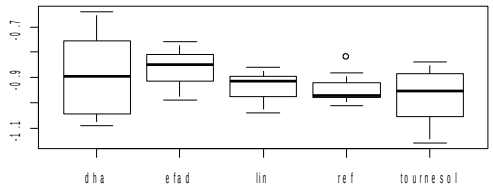
```
            Df  Sum Sq Mean Sq F value Pr(>F)
facteur      4 0.03353 0.00838  0.5652 0.6895
Residuals   35 0.51918 0.01483
```



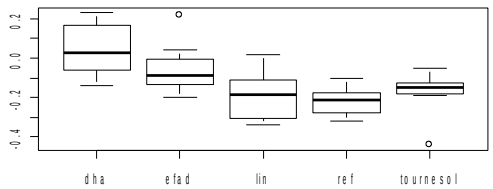

```
            Df  Sum Sq Mean Sq F value Pr(>F)
facteur      4 0.18796 0.04699  0.9322 0.4566
Residuals   35 1.76434 0.05041
```




```
            Df  Sum Sq Mean Sq F value Pr(>F)
facteur      4 0.06372 0.01593  1.4155 0.2492
Residuals   35 0.39391 0.01125
```




```
            Df  Sum Sq Mean Sq F value    Pr(>F)
facteur      4 0.39496 0.09874  7.0847 0.0002741 ***
Residuals   35 0.48780 0.01394
```



L'effet du facteur est significatif seulement dans le dernier cas.
Au moins 1 des 5 moyennes est différente des autres.

ANOVA 2 facteurs

Notations

Facteur A (I modalités)	Facteur B (J modalités)				Moyenne
	1	2	…	J	
1	Y_{111} … $\bar{Y}_{11.}$ … Y_{11K}	Y_{121} … $\bar{Y}_{12.}$ … Y_{12K}	…	Y_{1J1} … $\bar{Y}_{1J.}$ … Y_{1JK}	$\bar{Y}_{1..}$
2	Y_{211} … $\bar{Y}_{21.}$ … Y_{21K}	Y_{221} … $\bar{Y}_{22.}$ … Y_{22K}	…	Y_{2J1} … $\bar{Y}_{2J.}$ … Y_{2JK}	$\bar{Y}_{2..}$
…	…	…		…	
I	Y_{I11} … $\bar{Y}_{I1.}$ … Y_{I1K}	Y_{I21} … $\bar{Y}_{I2.}$ … Y_{I2K}	…	Y_{IJ1} … $\bar{Y}_{IJ.}$ … Y_{IJK}	$\bar{Y}_{I..}$
Moyenne	$\bar{Y}_{.1.}$	$\bar{Y}_{.2.}$	…	$\bar{Y}_{.J.}$	$\bar{Y}_{...}$

ANOVA 2 facteurs

Modèle et sommes de carrés

Modèle

$$Y_{ijk}=\mu+\alpha_i+\beta_j+\gamma_{ij}+\varepsilon_{ijk}$$

μ : effet moyen général

α_i : effet différentiel de la i^{ème} modalité du facteur A

β_i : effet différentiel de la j^{ème} modalité du facteur B

γ_{ij} : effet d'interaction de la i^{ème} modalité du facteur A avec la j^{ème} du facteur B

Décomposition

$$SC_{totale} \quad \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\left(y_{ijk}-\bar{y}_{...}\right)^2=$$

$$JK\sum_{i=1}^{I}\left(\bar{y}_{i..}-\bar{y}_{...}\right)^2+IK\sum_{i=1}^{I}\left(\bar{y}_{.j.}-\bar{y}_{...}\right)^2+K\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\bar{y}_{ij.}-\bar{y}_{i..}-\bar{y}_{.j.}+\bar{y}_{...}\right)^2+\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\left(y_{ijk}-\bar{y}_{ij.}\right)^2$$

$SC_{facteur\ A}$ $SC_{facteur\ B}$ $SC_{interaction}$ $SC_{résidu}$

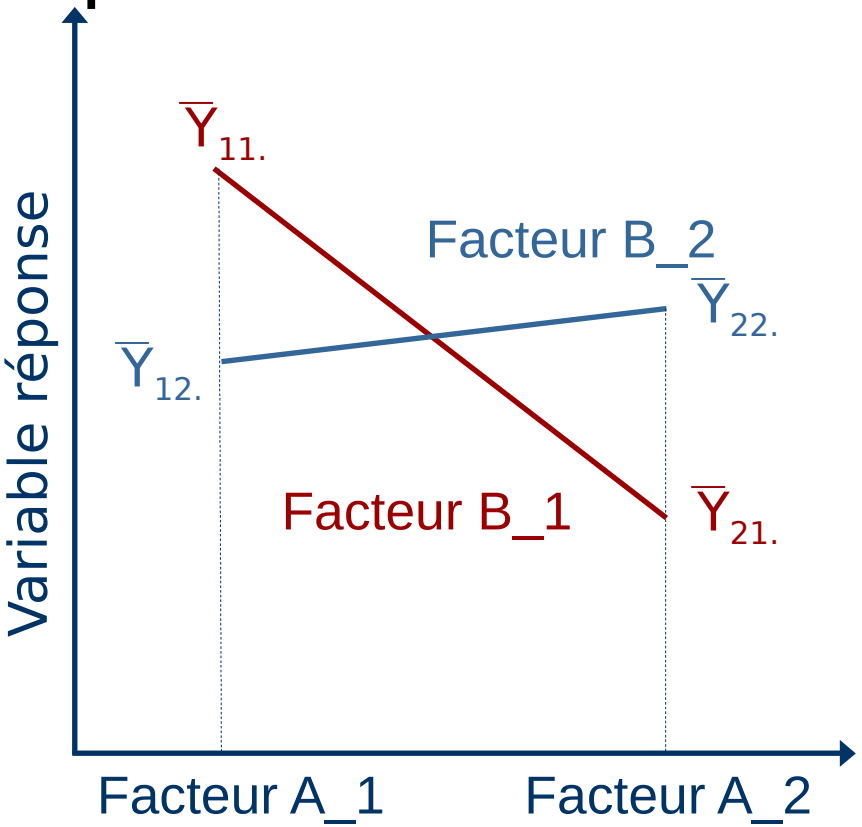
ANOVA 2 facteurs

Table d'ANOVA

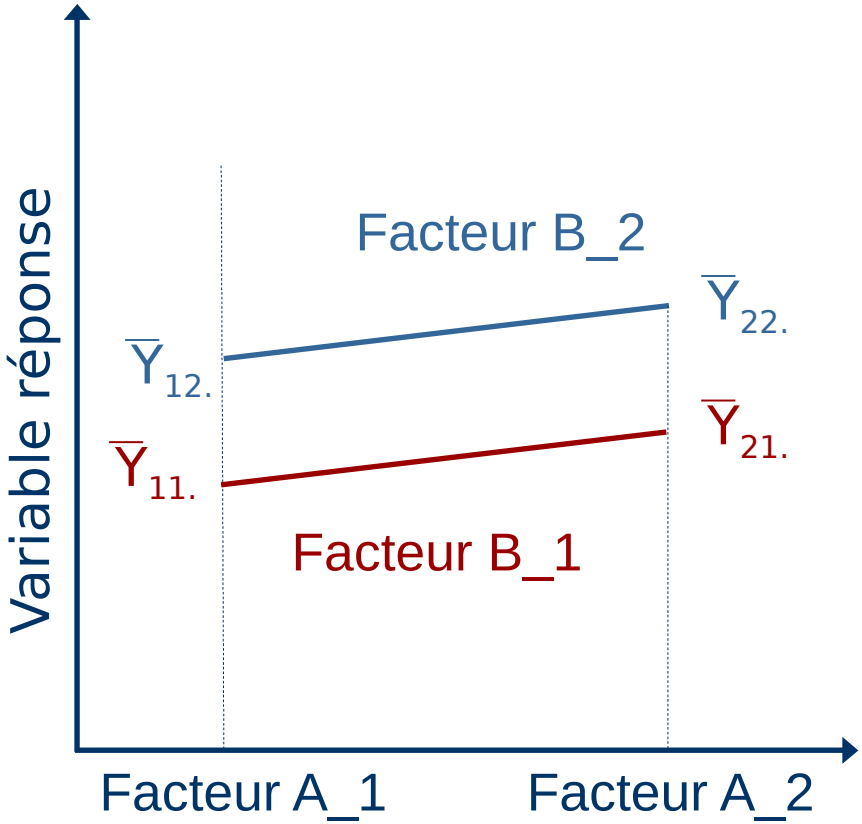
Sources de variation	Degrés de liberté	SCE (Somme de carrés d'écarts)	CM (carré moyen)	F	p-value
Facteur A	$I-1$	SCE_A	$CM_A = SCE_A/(I-1)$	$CM_A / CM_{résidu}$	...
Facteur B	$J-1$	SCE_B	$CM_B = SCE_B/(J-1)$	$CM_B / CM_{résidu}$	...
Interaction	$(I-1)(J-1)$	SCE_{A*B}	$CM_{A*B} = SCE_{A*B} / (I-1)(J-1)$	$CM_{A*B} / CM_{résidu}$	...
Résiduelle	$IJ(K-1)$	$SCE_{résidu}$	$CM_{resid} = SCE_{résidu}/IJ(K-1)$		

ANOVA 2 facteurs

Graphe d'interaction



L'effet d'un facteur dépend de la modalité de l'autre facteur ➔ **l'interaction** sera probablement déclarée **significative** lors du test statistique



L'effet d'un facteur ne dépend pas de la modalité de l'autre facteur ➔ **l'interaction** sera probablement déclarée **non significative** lors du test statistique

Modèle linéaire

$$Y = X\beta + \varepsilon$$

Régression linéaire

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{pmatrix} \quad X = \begin{pmatrix} X_1 & 1 \\ X_2 & 1 \\ \dots & \dots \\ X_n & 1 \end{pmatrix} \quad \beta = \begin{pmatrix} a \\ b \end{pmatrix} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{pmatrix}$$

ANOVA 1 facteur

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{pmatrix} \quad X = \begin{pmatrix} 1 & 0 \\ \dots & \dots \\ 1 & 0 \\ 0 & 1 \\ \dots & \dots \\ 0 & 1 \end{pmatrix} \quad \beta = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{pmatrix}$$

$$\hat{\beta} = (X'X)^{-1} X'Y$$

À suivre...