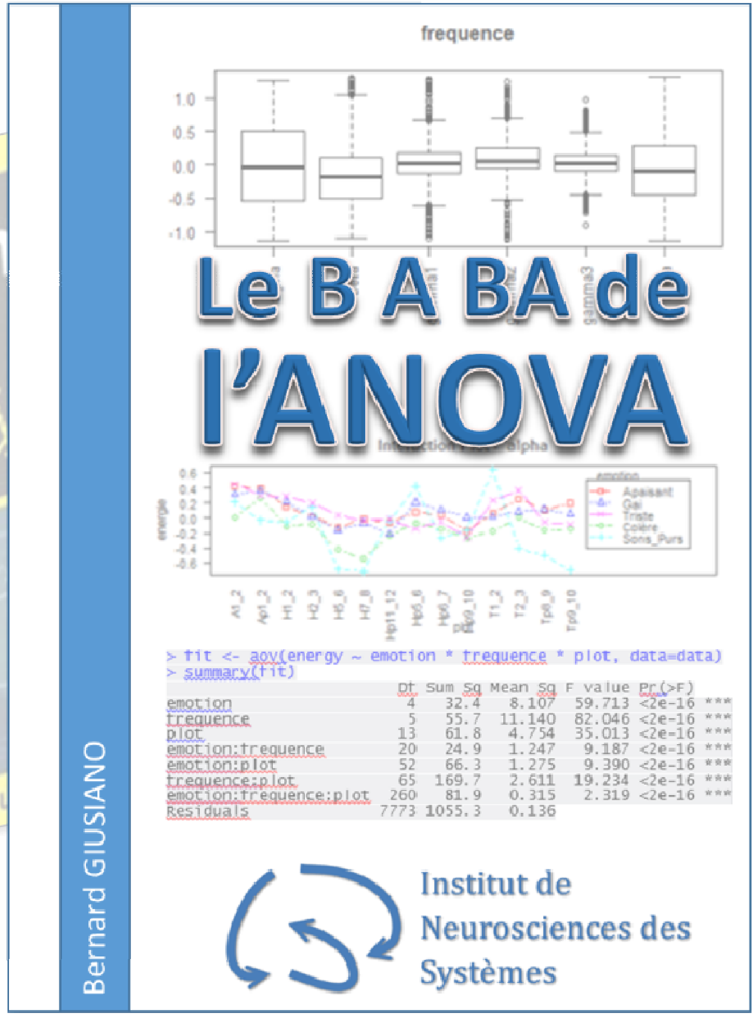
Qu’est-ce que c’est ?

A quoi ça sert ?

Pourquoi ne pas se satisfaire de simples tests de t 2 par 2 ?



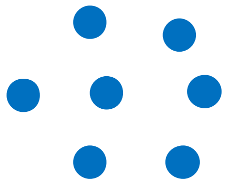
Trois individus (ou objets, ou sujets, …)

à chacun desquels on affecte une mesure

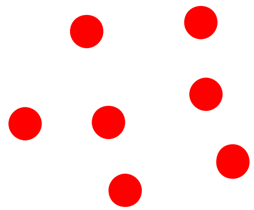
44 44 52

Bleu = Rouge Rouge ≠ Vert Bleu ≠ Vert

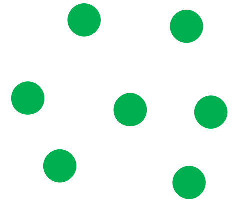
Trois groupes d’individus



34
43
66
44
28
56
23



57
80
20
37
27
41
32



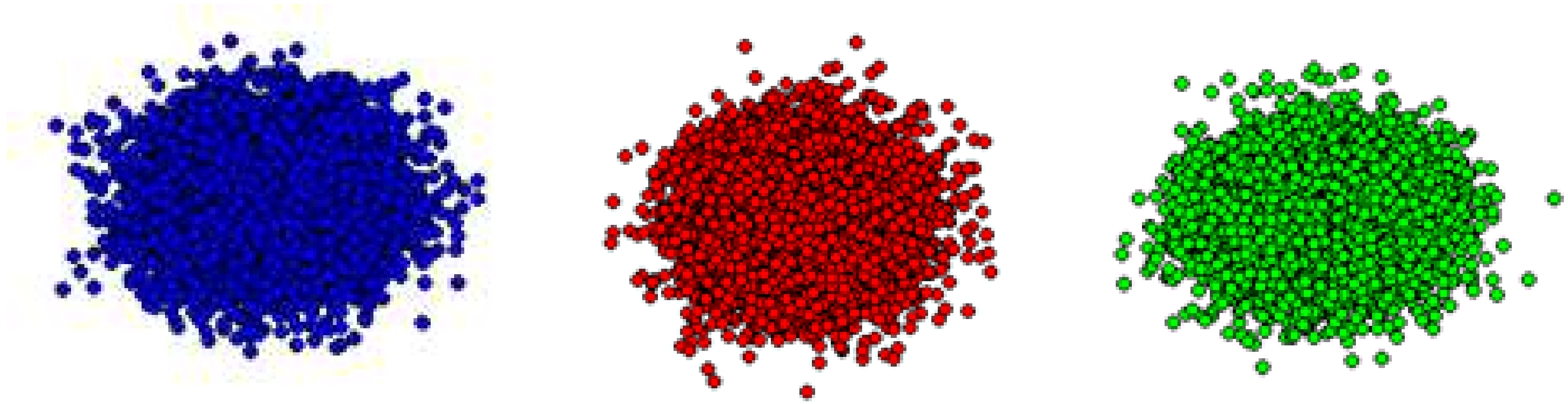
62
45
24
41
37
56
64

dont on calcule la moyenne des mesures

42 42 47

$m_{bleu} = m_{rouge}$ $m_{rouge} \neq m_{vert}$ $m_{bleu} \neq m_{vert}$

Trois très grands groupes d'individus
*(on parle de **populations**)*



pour lesquels on ne peut accéder aux mesures de tous les individus

$m_{bleu} \overset{?}{=} m_{rouge}$ $m_{rouge} \overset{?}{\neq} m_{vert}$ $m_{bleu} \overset{?}{\neq} m_{vert}$

On peut démontrer que si on tire un **échantillon** représentatif de la population *(par tirage au hasard de 30 individus, par exemple)*,

la moyenne de cette échantillon ($\bar{x}$) est une bonne **estimation** ($\hat{\mu}$) de la moyenne de la population (μ) :

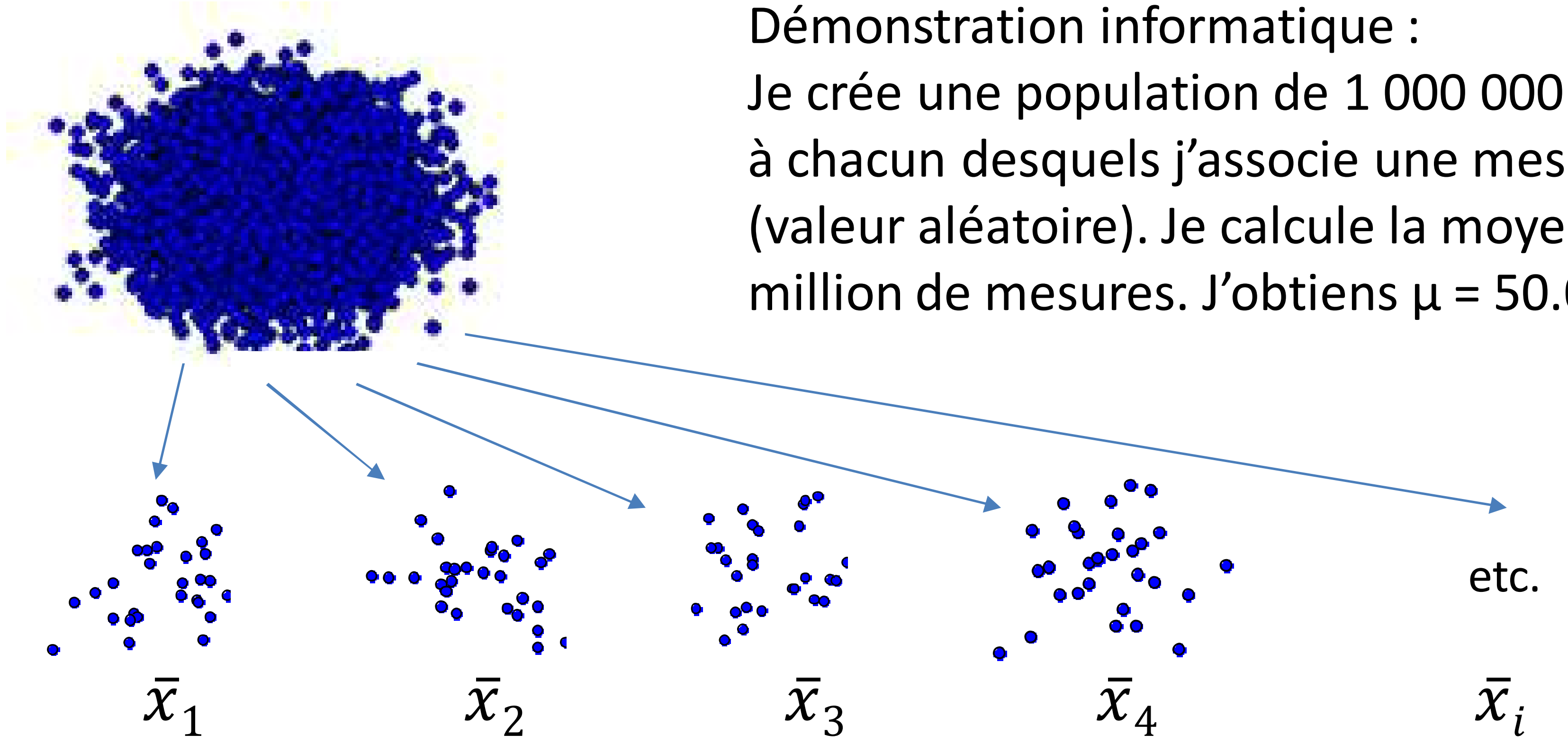
$$\hat{\mu} = \bar{x}$$

Bonne estimation : l’espérance mathématique de $\hat{\mu}$ est égale à μ

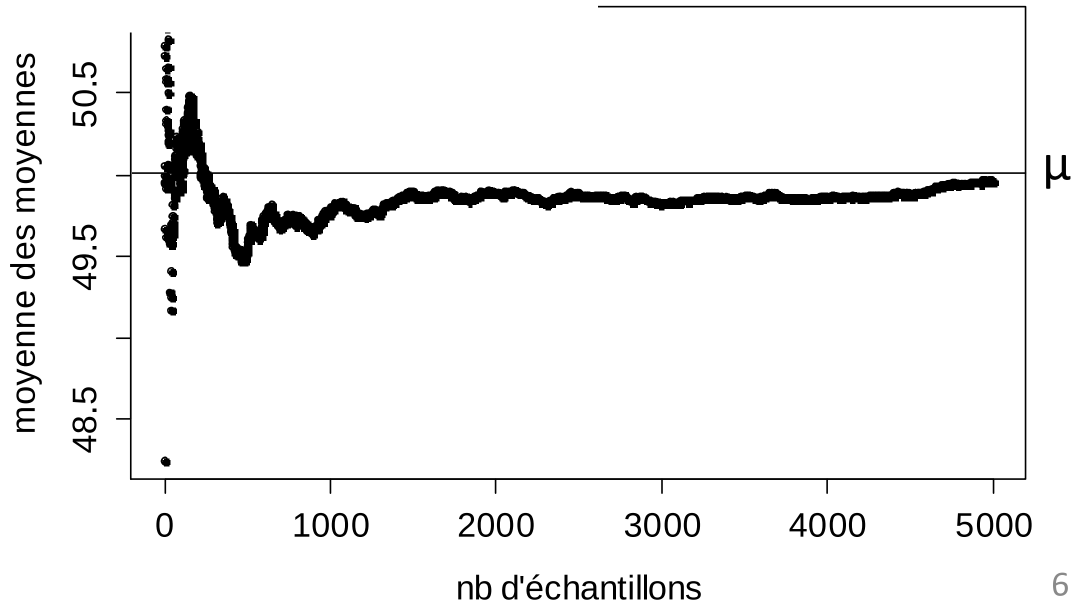
$$E(\hat{\mu}) = \mu$$

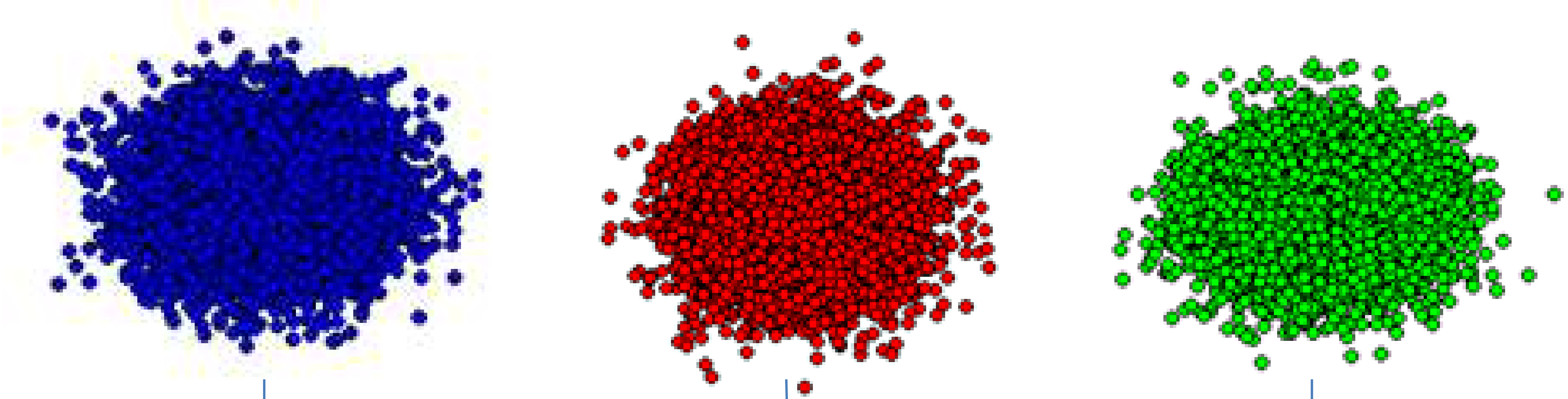
Ce qui signifie que si on tire un nombre très grand d’échantillons de même effectif, la moyenne *(espérance)* des moyennes $\bar{x}$ de ces échantillons tend vers μ, moyenne de la population.

Démonstration informatique :
Je crée une population de 1 000 000 individus à chacun desquels j'associe une mesure (valeur aléatoire). Je calcule la moyenne de ce million de mesures. J'obtiens μ = 50.01944

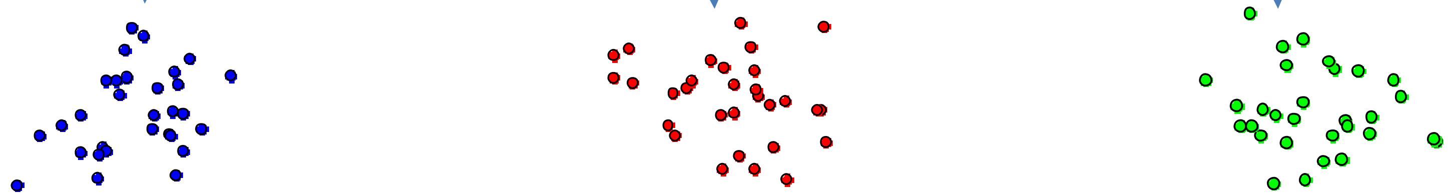


Je tire i échantillons de 30 individus chacun.
Je calcule leur moyenne.
Quand i tend vers l'infini, la moyenne des $\bar{x}_i$ tend vers μ.





Tirage au hasard de 30 individus dans chacune des 3 populations



$\bar{x}_b = 3.93$ $\neq$ $\bar{x}_r = 4.28$ $\neq$ $\bar{x}_v = 5.17$

$\hat{\mu}_b \neq \hat{\mu}_r$ $\hat{\mu}_r \neq \hat{\mu}_v$ $\hat{\mu}_b \neq \hat{\mu}_v$

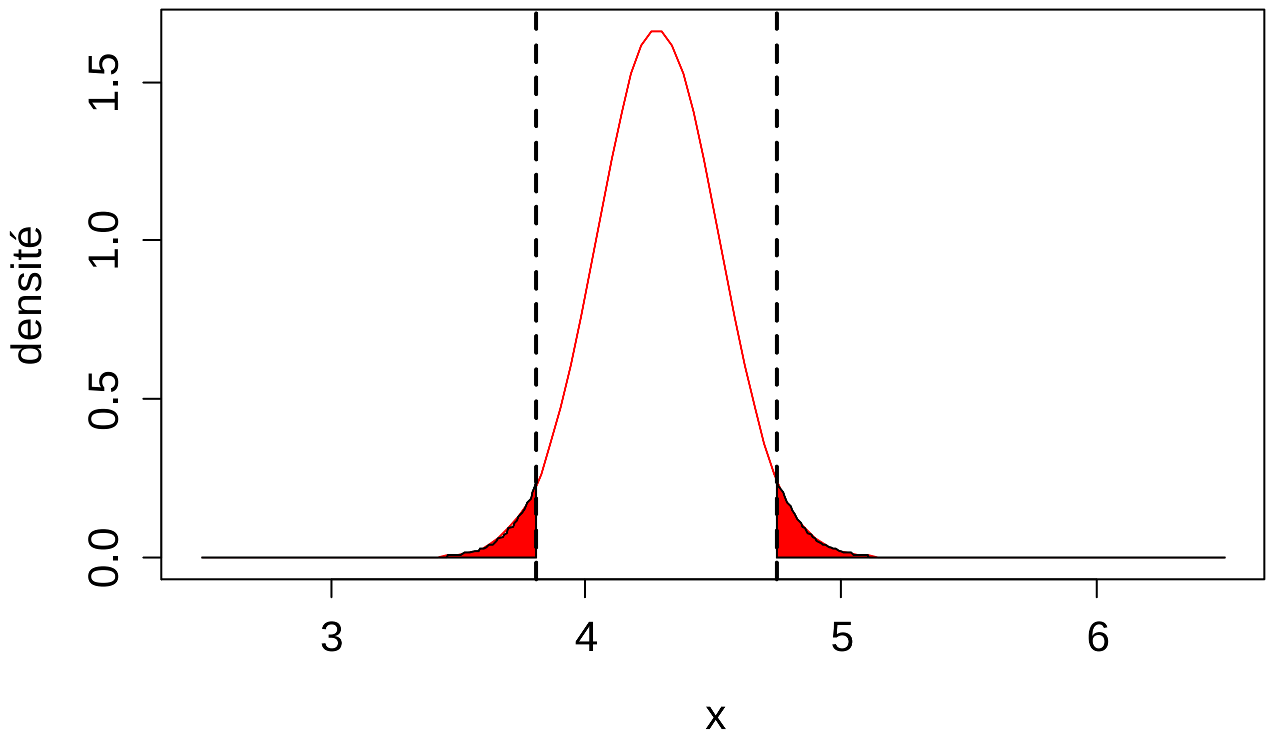
Peut-on en conclure que les moyennes des 3 populations sont différentes ?

$\mu_b \overset{?}{\neq} \mu_r$ $\mu_r \overset{?}{\neq} \mu_v$ $\mu_b \overset{?}{\neq} \mu_v$

NON, car $\hat{\mu}$ n'est pas μ, ce n'en est qu'une **estimation**.

Sa valeur doit être accompagnée de son **intervalle de confiance** qui donne la précision de l'estimation de la moyenne de la population à partir de l'échantillon.

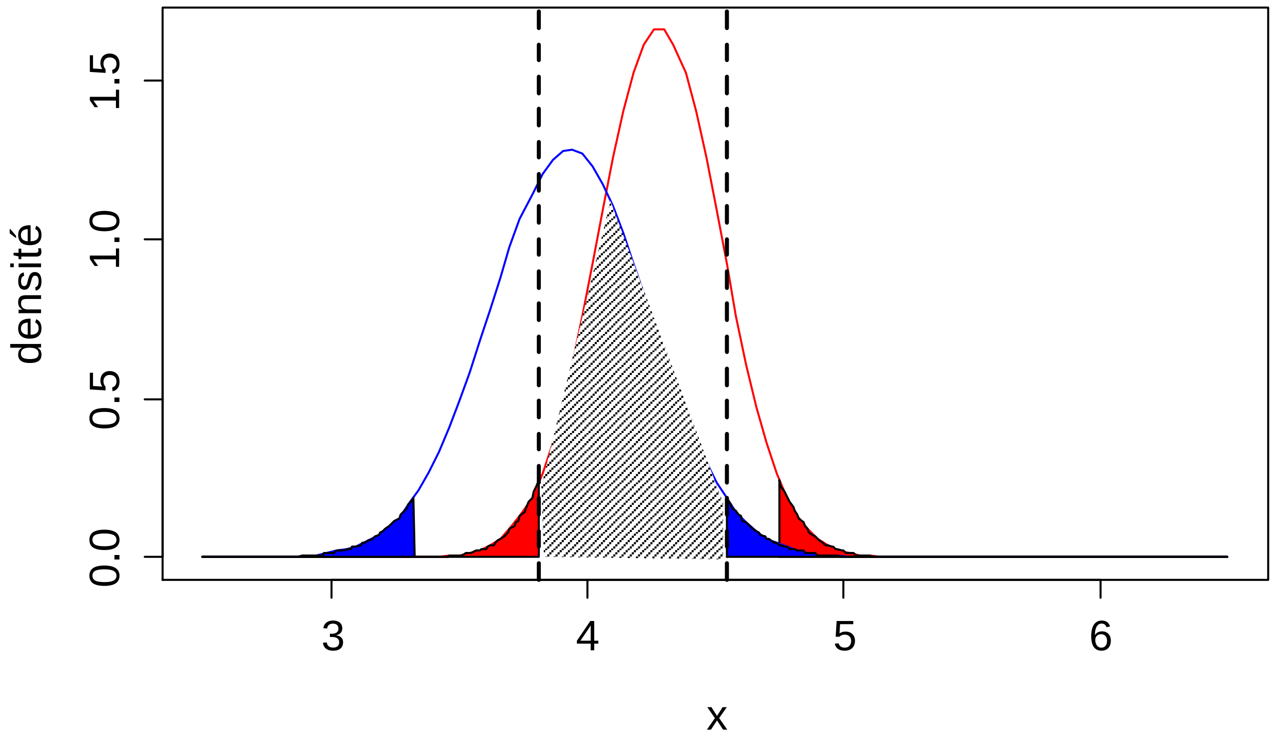
Si on répétait l'estimation un grand nombre de fois, dans (1-α) % des cas l'intervalle de confiance contiendrait la vraie moyenne de la population, μ.



Dans cette estimation à partir d'un échantillon de la population rouge, la moyenne μ pourrait bien être dans l'intervalle **[3.81 ; 4.75]** avec un **degré de confiance** de 95%.

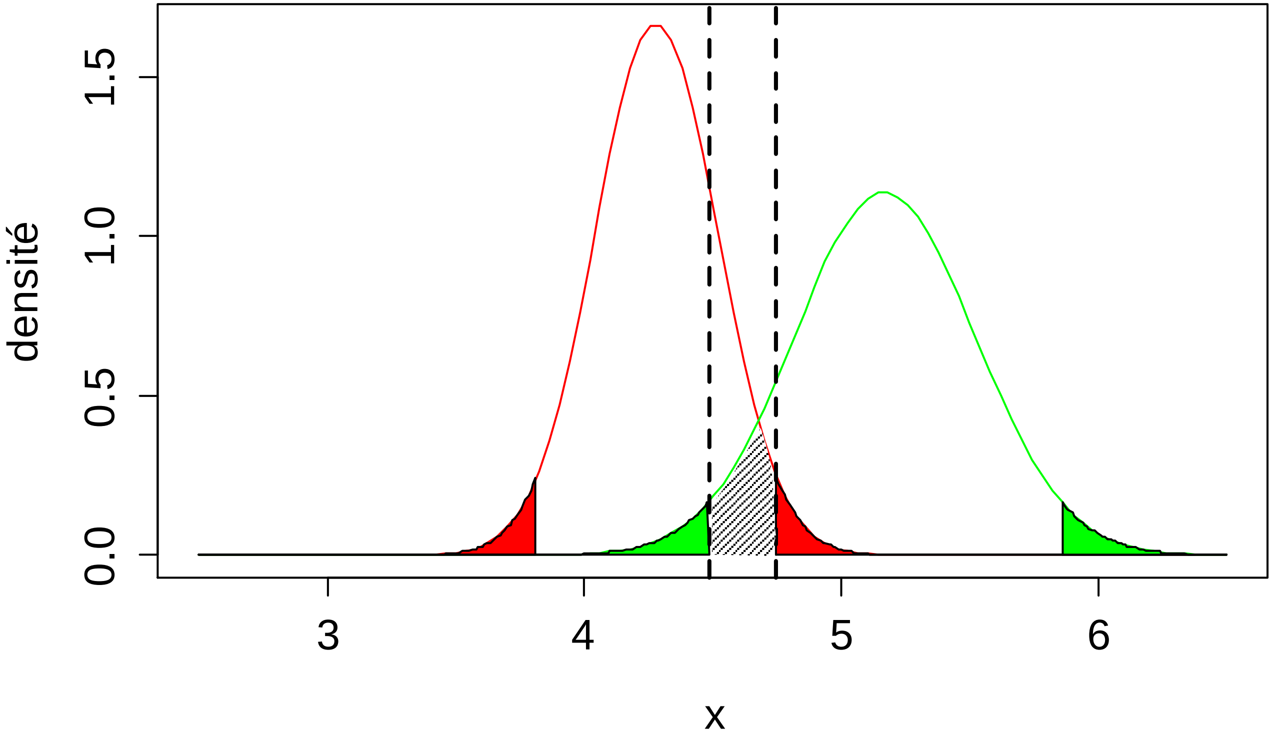
$\hat{\mu}_r$ et $\hat{\mu}_b$ ne sont que des **estimations**.

Si la population rouge et la population bleue ont la même moyenne, µ, celle-ci pourrait bien être dans l'intervalle **[3.81 ; 4.54]** avec un **degré de confiance** équivalent à l'aire hachurée.

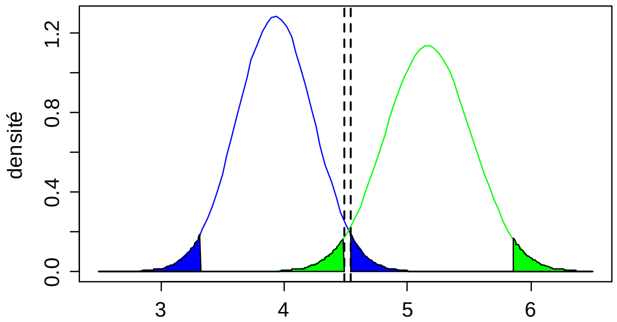


$\hat{\mu}_r$ et $\hat{\mu}_v$ ne sont que des **estimations**.

Si la population rouge et la population verte ont la même moyenne, µ, celle-ci pourrait bien être dans l'intervalle **[4.48 ; 4.75]** avec un **degré de confiance** équivalent à l'aire hachurée.

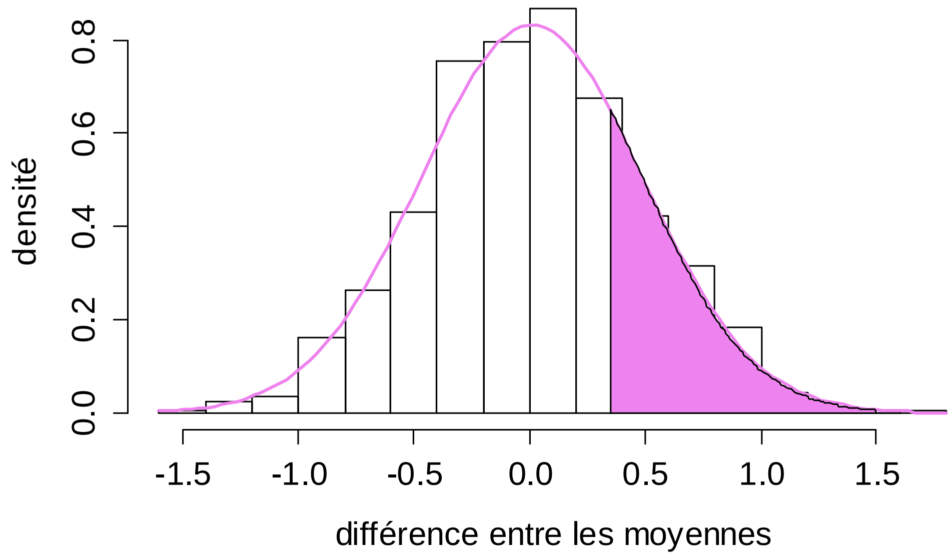


On comprend que si le degré de confiance est *« trop petit »*, on choisira l'**hypothèse** la plus vraisemblable : les deux populations n'ont pas la même moyenne.



Faisons l'**hypothèse** que $\mu_r = \mu_b$ alors $\mu_r - \mu_b = 0$

Si nous répétons un grand nombre de fois le tirage d'un échantillon dans chacune de deux populations de même moyenne $\mu = \mu_r = \mu_b$, que nous calculons chaque fois la différence entre les moyennes des deux échantillons, **cette différence suit une loi normale de moyenne μ = 0**.



Reprenons nos échantillons rouge et bleu :

$\bar{x}_r = 4.28$ et $\bar{x}_b = 3.93$

donc $\bar{x}_r - \bar{x}_b = 0.35$

La probabilité que la différence de moyennes entre les deux échantillons soit égale ou supérieure à 0.35, alors que les populations dont ils sont tirés ont des moyennes égales ***(hypothèse nulle)***, est donnée par l'aire colorée en violet. C'est le « petit p ».

Conclusion de ce rappel

- Si la question est :
  « les moyennes des **échantillons** sont-elles différentes ? »
  nous n'avons pas besoin d'un test d'hypothèse.

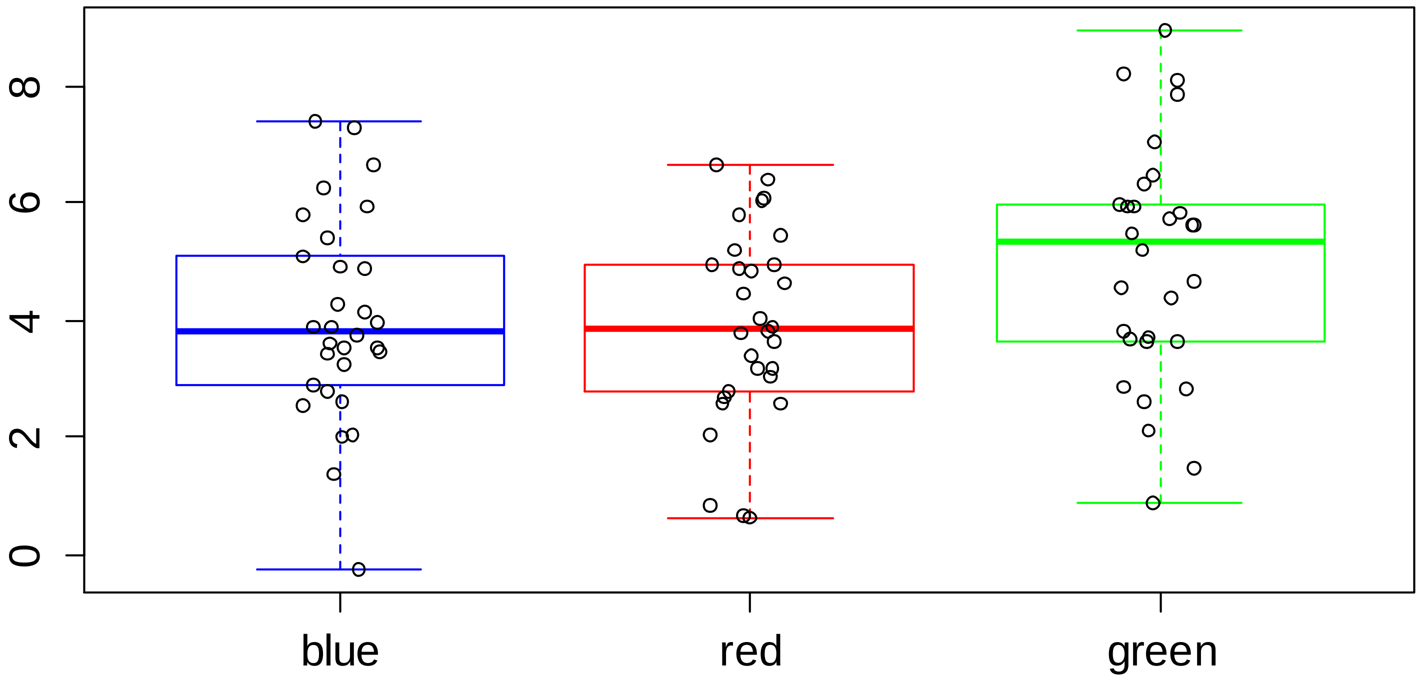
  Exemple : Est-ce que la mesure moyenne est différente entre mes 6 patients et mes 8 sujets contrôle ?

- Si la question est :
  « au vu des échantillons, est-ce que les moyennes des **populations** dont ils sont tirés sont différentes ? »
  nous avons besoin d'un **test d'hypothèse**…
  … et nous prenons des risques
  liés à l'incertitude sur les estimations.

  Exemple : A partir de mes 6 patients et de mes 8 sujets contrôle, est-il raisonnable de dire qu'en général la mesure moyenne des patients est différente de celle des sujets contrôle ?

- L'incertitude est contrôlée par les **conditions d'application** des tests.

ANOVA à 1 facteur (ayant 3 modalités) - 1 way ANOVA (3 levels)



Nombre de groupes : k = 3 {bleu, rouge, vert}

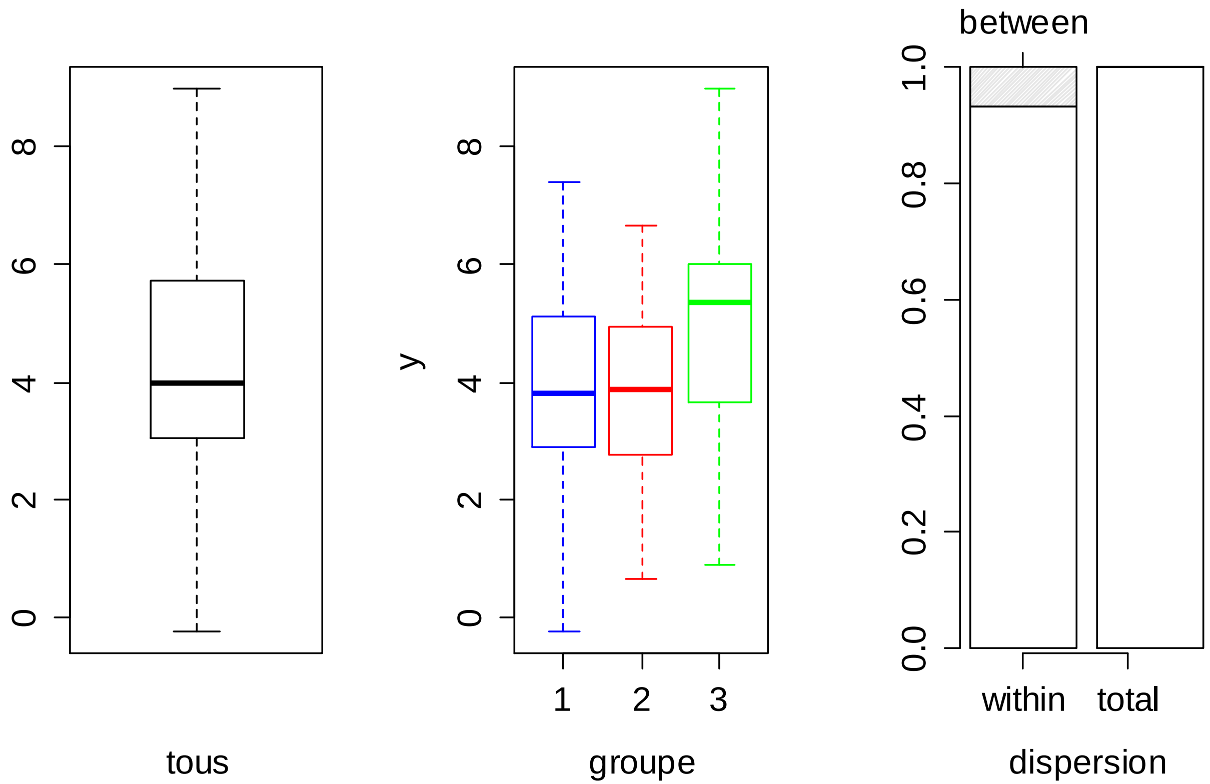
Effectif par groupe : $n_{bleu} = n_{rouge} = n_{vert} = 30$

Effectif total : n = 90

L'analyse de la variance   ANalysis Of VAriance (ANOVA)

Elle repose sur la décomposition de la variance totale :

$$\text{variance}_{\text{totale}} \equiv \text{variance}_{\text{entre les groupes}} + \text{variance}_{\text{à l'intérieur des groupes}}$$



Mais il s'agit bien de comparer des moyennes :

$\text{variance}_{\text{totale}}$: différence entre valeurs et moyenne globale

$$SCE^{*}_{totale} = \sum_{i}^{n} (x_i - \bar{x})^2 = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x})^2$$

$\text{variance}_{\text{entre les groupes}}$: différence entre moyennes des groupes et moyenne globale

$$SCE_{inter\ groupe} = \sum_{j=1}^{k} n_j (\bar{x}_j - \bar{x})^2 = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (\bar{x}_j - \bar{x})^2$$

$\text{variance}_{\text{à l'intérieur des groupes}}$: différence entre valeurs et moyennes des groupes

$$SCE_{intra\ groupe} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2$$

**SCE : Somme des Carrés des Ecarts à la moyenne*

La variance est la moyenne des carrés des écarts :

estimation de la variance$_{\text{totale}}$ = $s^2_{totale} = \frac{SCE_{totale}}{ddl^*_{total}} = \frac{SCE_{totale}}{n-1}$

estimation de la variance$_{\text{entre les groupes}}$ = $s^2_{intra} = \frac{SCE_{inter}}{ddl_{inter}} = \frac{SCE_{inter}}{k-1}$

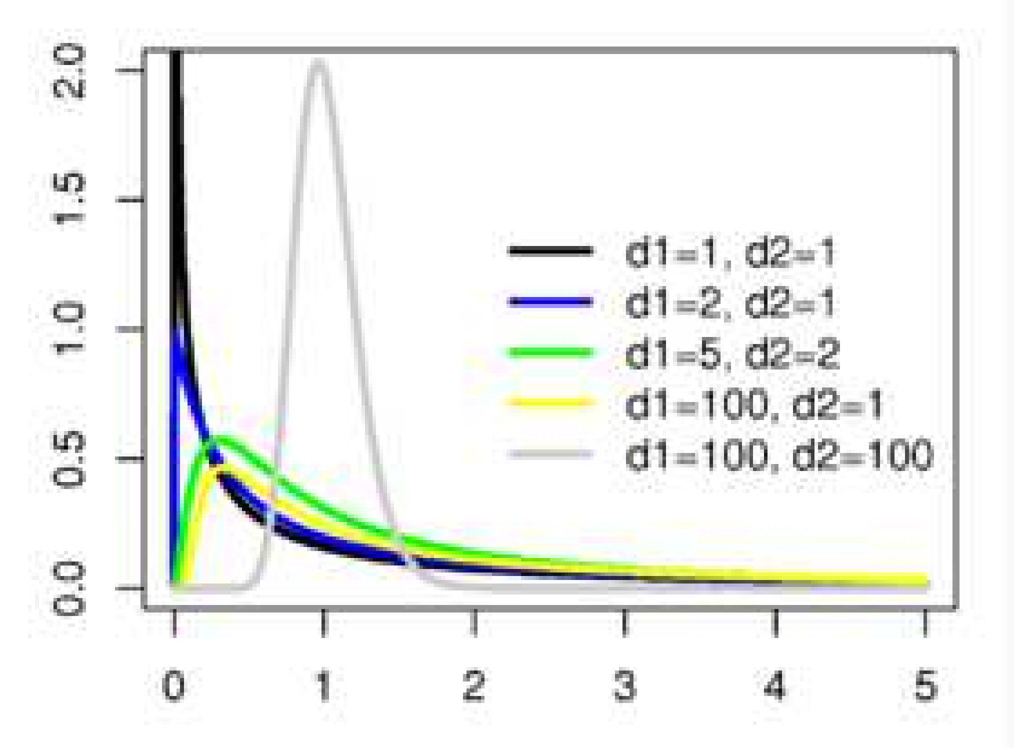
estimation de la variance$_{\text{à l'intérieur des groupes}}$ = $s^2_{inter} = \frac{SCE_{intra}}{ddl_{intra}} = \frac{SCE_{intra}}{n-k}$

$$SCE_{totale} = SCE_{inter} + SCE_{intra}$$

**ddl : nombre de degrés de liberté*

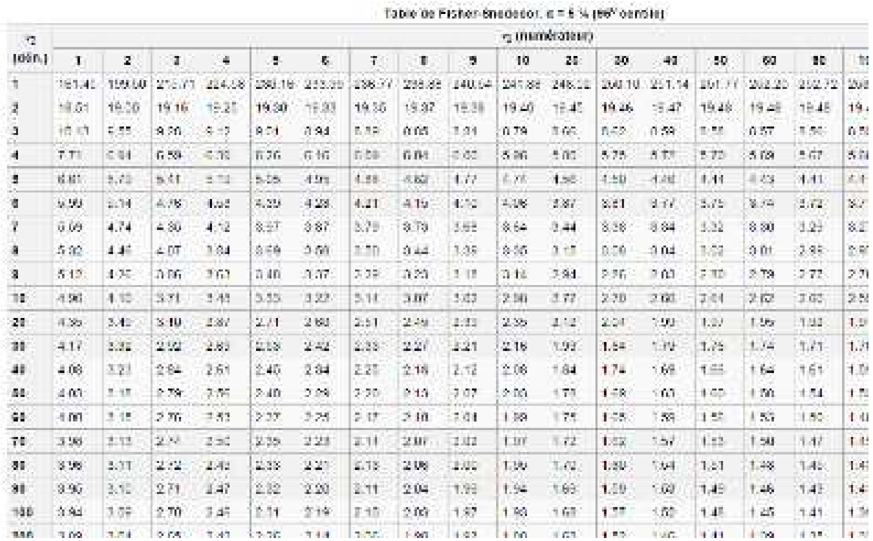
Dans l'hypothèse nulle (H_0) où il n'y a pas de différence entre les moyennes des 3 populations d'où sont tirés les 3 échantillons,

le rapport entre la variance inter et la variance intra suit une loi de distribution connue, la **loi du F de Fisher-Snedecor**,
à $\nu_1 = k - 1$ et $\nu_2 = n - k$ degrés de liberté.



Densité de probabilité (ou fonction de masse)

$$F = \frac{s^2_{inter}}{s^2_{intra}} = \frac{s^2_{factorielle}}{s^2_{résiduelle}}$$



ANOVA des données de notre exemple :

le facteur « groupe » {bleu, rouge, vert}

nombre de degrés de liberté

somme des carrés des écarts

moyenne des carrés des écarts

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
factor	2	20.78	10.392	3.171	0.0469	*
Residuals	87	285.13	3.277			

les résidus (variance intra groupe)

valeur du F

La probabilité que, sous l'hypothèse nulle, la valeur du F soit ≥ 3.171 est de 0.0469 (donc **p < 0.05**).

Conclusion : rejet de l'hypothèse nulle *(au risque α)*, trop improbable. Les moyennes des trois groupes sont significativement différentes.

Conditions d'application de l'ANOVA

1. **Indépendance** des échantillons
   – La meilleure solution : tirer les échantillons au hasard.
2. **Normalité** des populations.
   – Il faut pouvoir considérer que les populations d'où sont tirés les échantillons sont distribuées selon la loi normale.
3. Egalité des variances (**homoscédasticité**)
   – Le facteur étudié peut influencer les moyennes, mais pas les variances.

Des méthodes existent pour vérifier ces 3 conditions d'application.

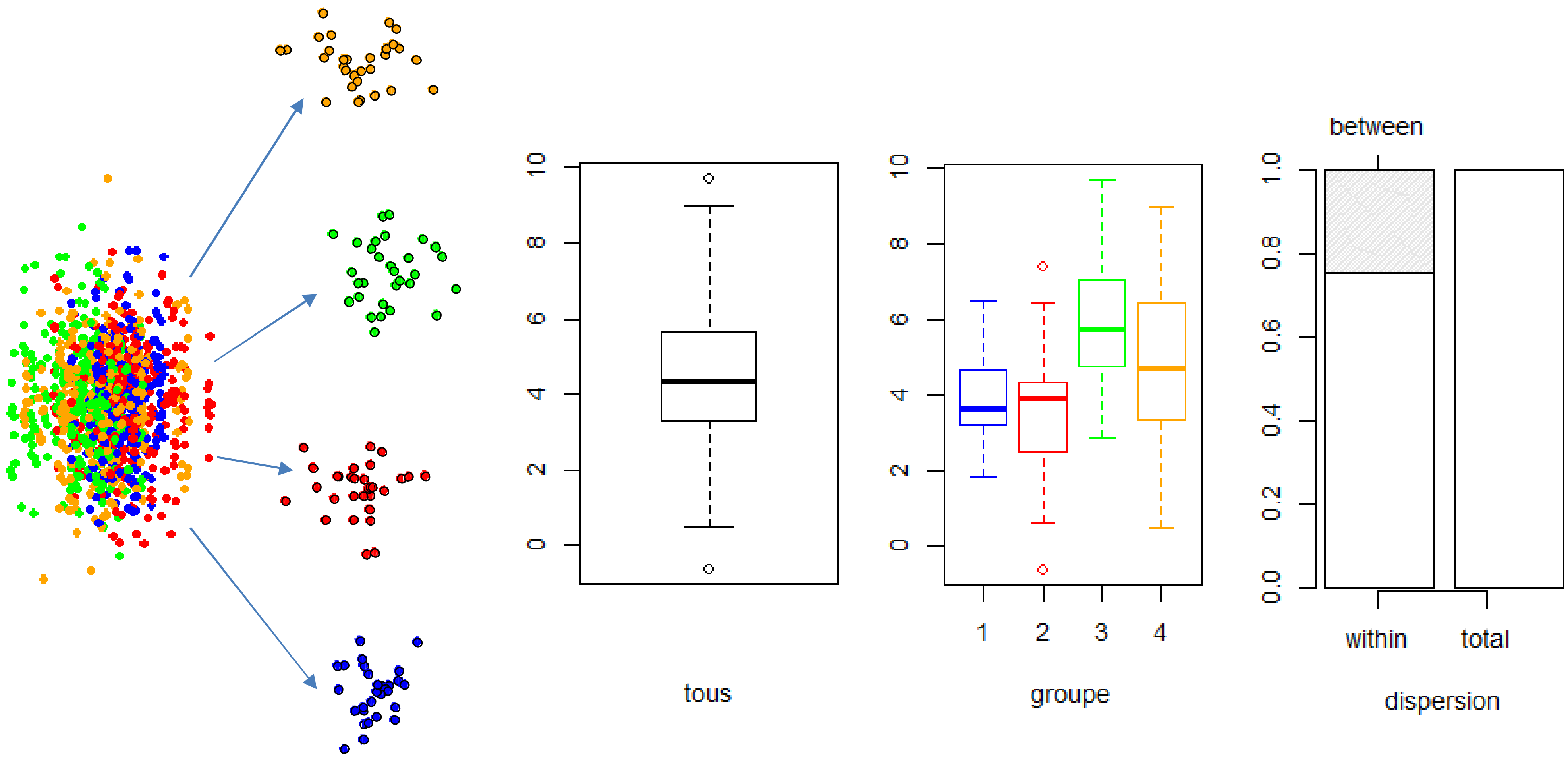
L'indépendance est la condition la plus critique.

L'ANOVA est robuste vis-à-vis de la normalité stricte et de l'homoscédasticité si les échantillons sont assez grands (> 30).

Si ces conditions ne peuvent pas être satisfaites, on peut utiliser le **test de Kruskal–Wallis** comme équivalent non paramétrique de l'analyse de variance à un facteur.

Mais quel groupe est différent de quel autre groupe ?

On utilise un <u>autre exemple</u> : une population dont on tire quatre échantillons indépendants et de même effectif (n_j = 30), un pour chaque couleur.



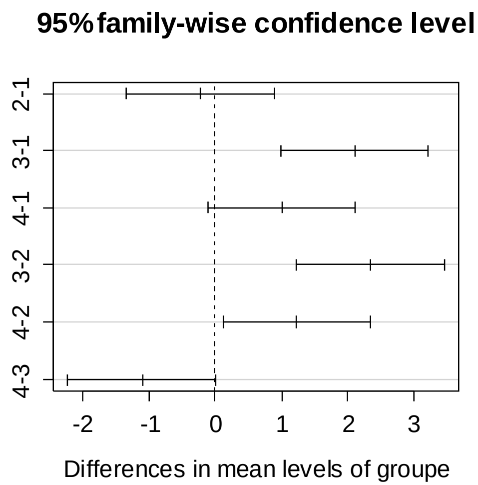
ANOVA de ce nouvel exemple :

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
factor	3	102.8	34.28	12.5	**3.86e-07**	***
Residuals	116	318.2	2.74			

Tests post-hoc de Tukey (Tukey Honest Significant Differences) :

$factor

	diff	lwr	upr	p adj
2-1	-0.2301661	-1.3448419	0.88450977	0.9495035
3-1	2.1078624	0.9931865	3.22253825	**0.0000164**
4-1	1.0041868	-0.1104890	2.11886266	0.0932113
3-2	2.3380285	1.2233526	3.45270433	**0.0000016**
4-2	1.2343529	0.1196770	2.34902874	**0.0237150**
4-3	-1.1036756	-2.2183514	0.01100026	0.0533531



La méthode de Tukey consiste à déterminer la différence minimum entre deux groupes qui peut être considérée comme significative.

L'ajustement de p pour les comparaisons multiples (ici au nombre de 6) est assuré par la référence à la distribution des intervalles de Student.

cf. Wikipedia : Tukey's range test

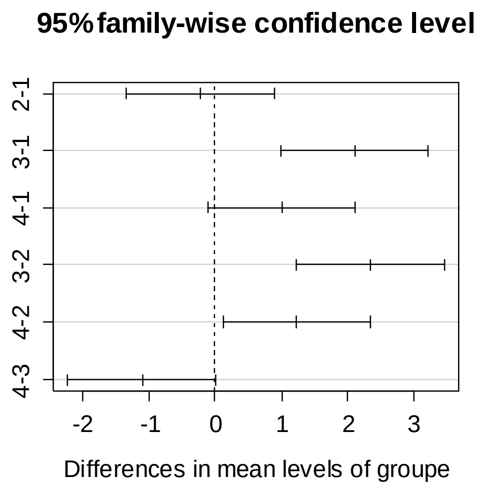
Quelle différence avec des test de t deux à deux ?

Si on réalise les 6 tests de comparaisons 2 à 2 avec le test du t de Student, se pose le problème des comparaisons multiples :

$\alpha_{global} \approx \alpha_{chaque\ test}$ x nb de comparaisons => $\alpha_{global} \approx 0.05 \times 6 = 0.30$

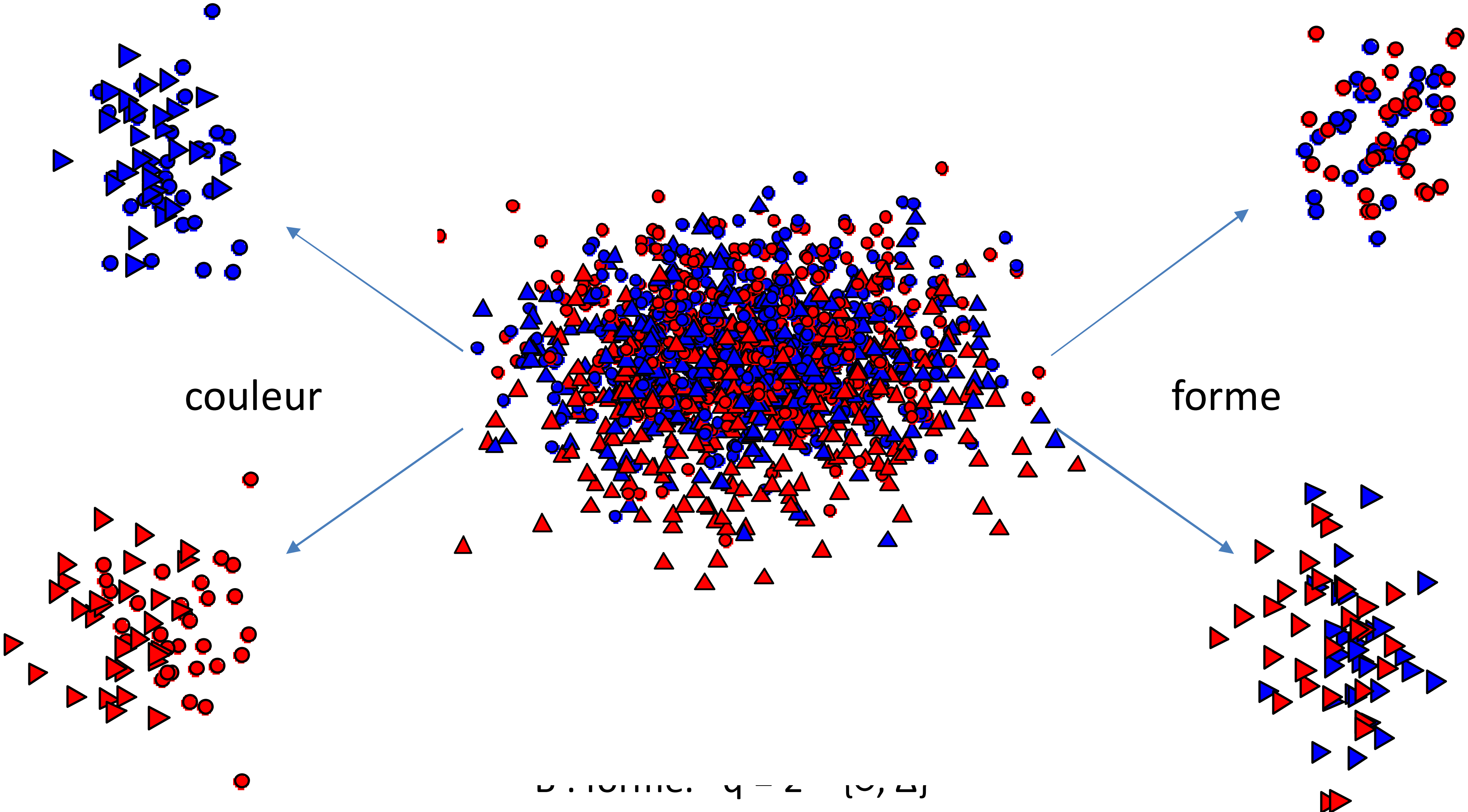
Traitons-le avec la correction de Bonferroni : $\alpha_{corrigé} = \alpha$/nb de comparaisons

	difference	**Tukey p adj**	test t p	**test t p adj**
2-1	-0.2301661	0.9495035	0.5417	1.0
3-1	2.1078624	**0.0000164**	0.0000006	**0.0000039**
4-1	1.0041868	0.0932113	0.02006	0.1203465
3-2	2.3380285	**0.0000016**	0.0000016	**0.0000098**
4-2	1.2343529	**0.0237150**	0.01237	**0.0742211**
4-3	-1.1036756	0.0533531	0.0237	0.1422164



Le test de t n'a pas mis en évidence la différence entre 4 et 2.
Le test de Tukey est donc plus **puissant** que le test de t corrigé.
Il est aussi plus **robuste** car il utilise toutes les données de l'expérience chaque fois, et pas seulement les données des deux groupes testés.

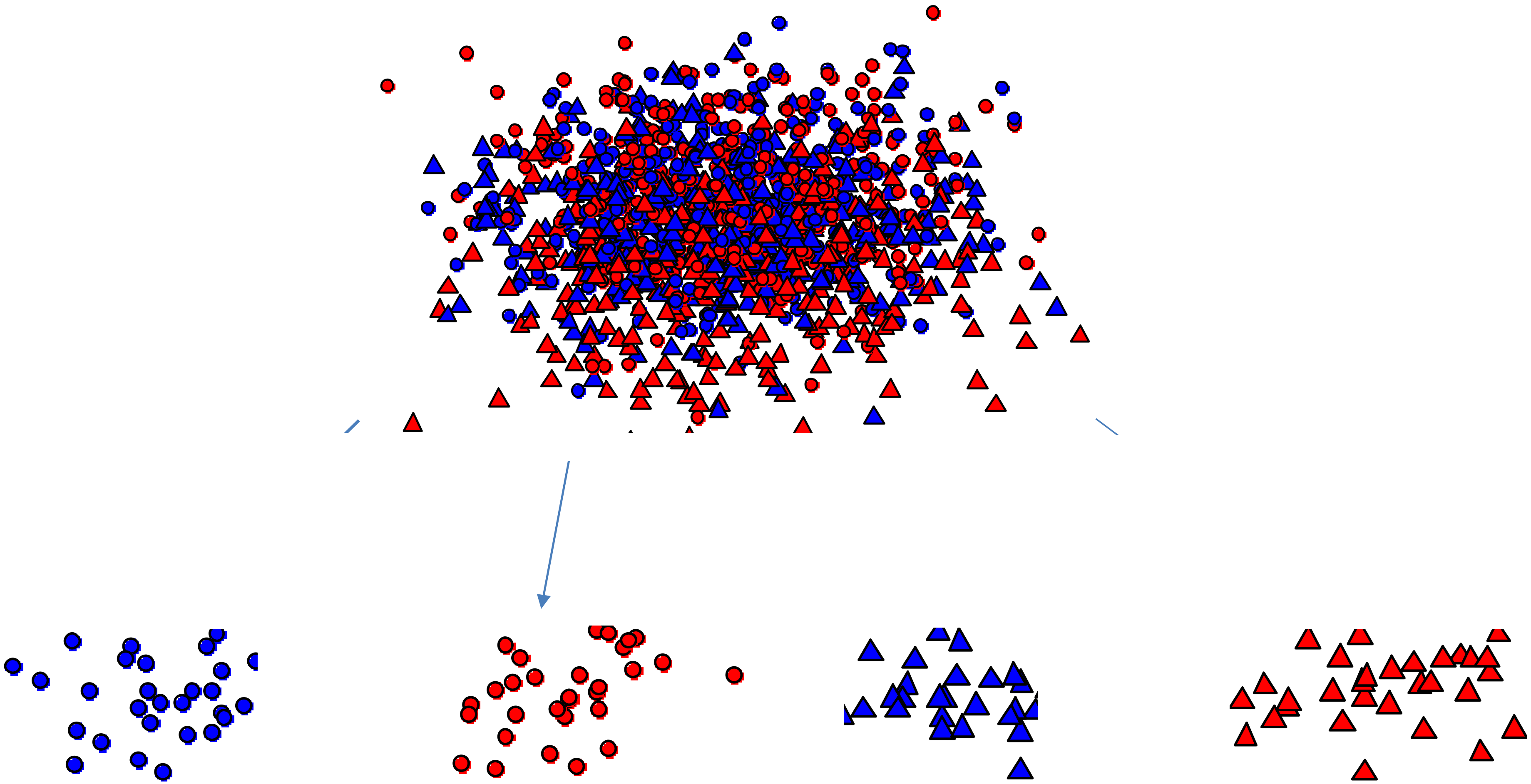
e l’ANOVA : l’analyse multifactorielle



Effectif par modalité des facteurs = 60

Effectif total : n = 120

ANOVA à 2 facteurs (2 modalités) - 2 ways ANOVA (2 levels)



ANOVA à 2 facteurs (2 modalités) - 2 ways ANOVA (2 levels)

$$SCE_{totale} = SCE_{inter} + SCE_{intra}$$

$$SCE_{Totale} = SCE_{Factorielle} + SCE_{Résiduelle}$$

$$SCE_{Factorielle} = SCE_{Facteur\ A} + SCE_{Facteur\ B} + SCE_{Interaction\ AB}$$

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
A	1	2.10	2.10	2.537	0.11390	
B	1	32.46	32.46	39.243	**6.56e-09**	***
A:B	1	6.85	6.85	8.284	**0.00476**	**
Residuals	116	95.96	0.83			

ANOVA à 2 facteurs (2 modalités) - 2 ways ANOVA (2 levels)

Error Decomposition

$$\underbrace{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}\left(Y_{ijk}-\bar{Y}_{...}\right)^2}_{SS_{Total}} = \underbrace{r\cdot b\cdot\sum_{i=1}^{a}\left(\bar{Y}_{i..}-\bar{Y}_{...}\right)^2}_{SS_A} + \underbrace{r\cdot a\cdot\sum_{j=1}^{3}\left(\bar{Y}_{.j.}-\bar{Y}_{...}\right)^2}_{SS_B}$$

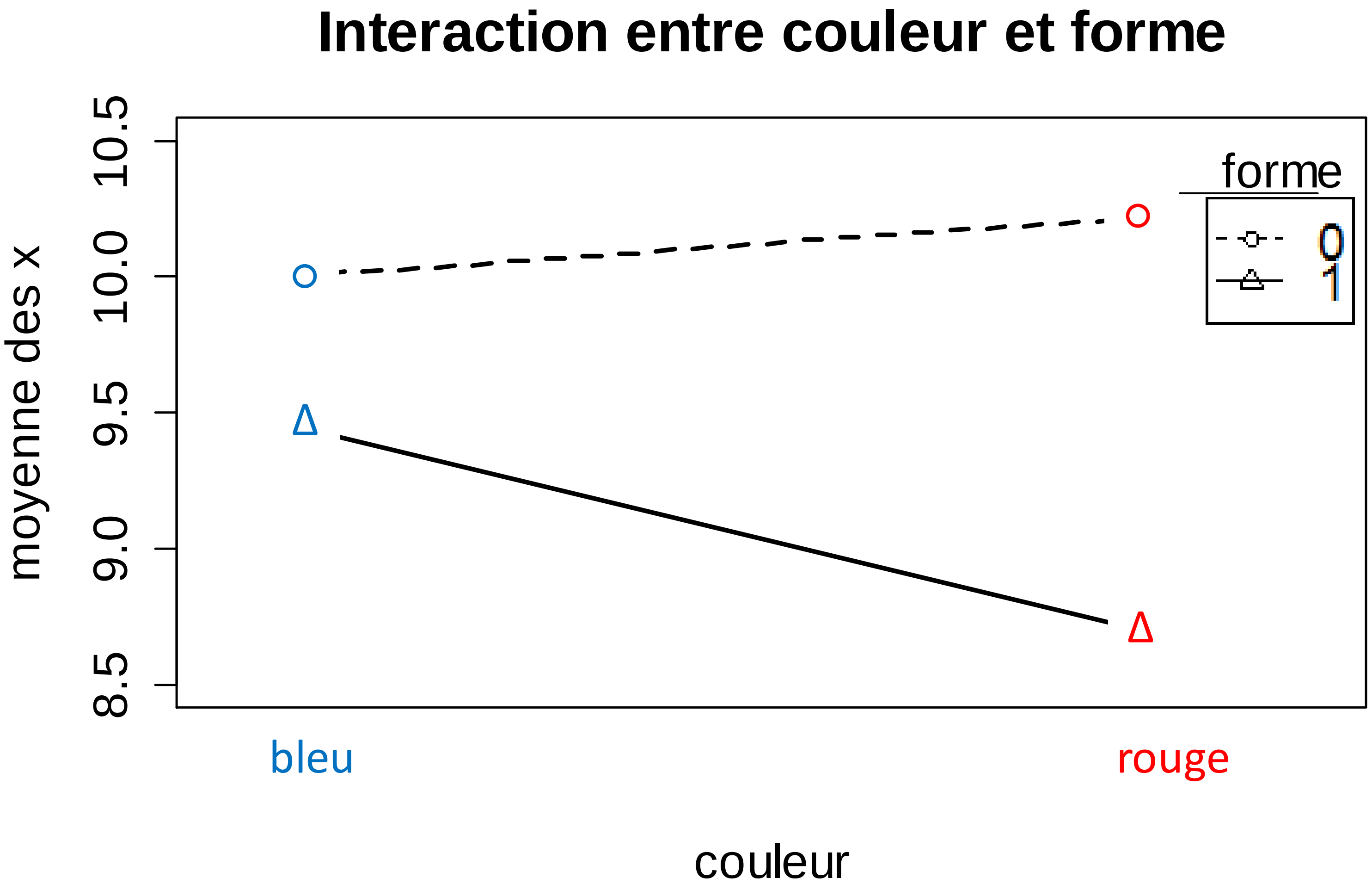
$$+\underbrace{r\times\sum_{i=1}^{a}\sum_{j=1}^{b}\left(\bar{Y}_{ij.}-\bar{Y}_{i..}-\bar{Y}_{.j.}+\bar{Y}_{...}\right)^2}_{SS_{A\times B}} + \underbrace{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}\left(Y_{ijk}-\bar{Y}_{ij.}\right)^2}_{SS_{within}}$$

ANOVA Table

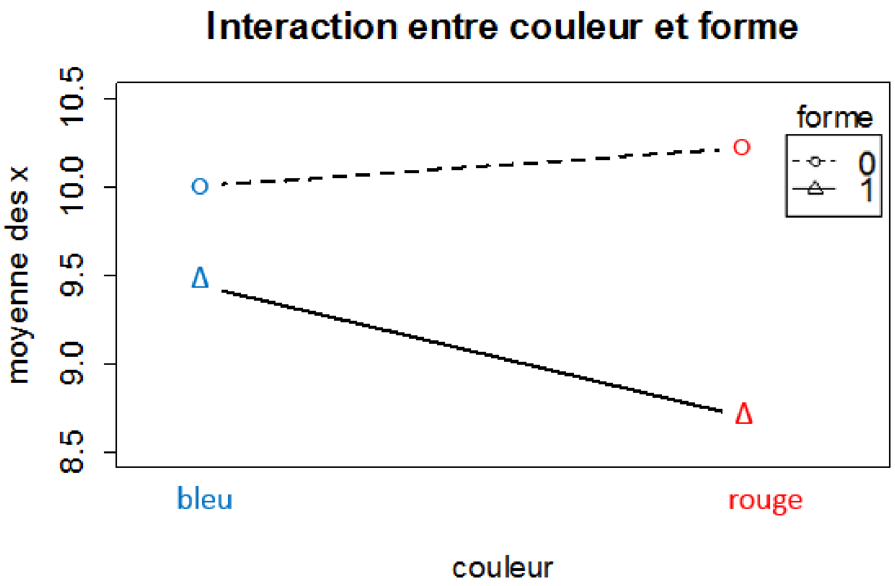
Source	Degrees of Freedom	SS	MS	F
A	a-1	SS_A	MS_A	MS_A / MS_{within}
B	b-1	SS_B	MS_B	MS_B / MS_{within}
$A\times B$	(a-1)(b-1)	$SS_{A\times B}$	$MS_{A\times B}$	$MS_{A\times B} / MS_{within}$
Within	ab(r-1)	SS_{within}	MS_{within}	
Total	abr-1	SS_{Total}		

Interaction : l'effet d'un facteur dépend du niveau de l'autre facteur.

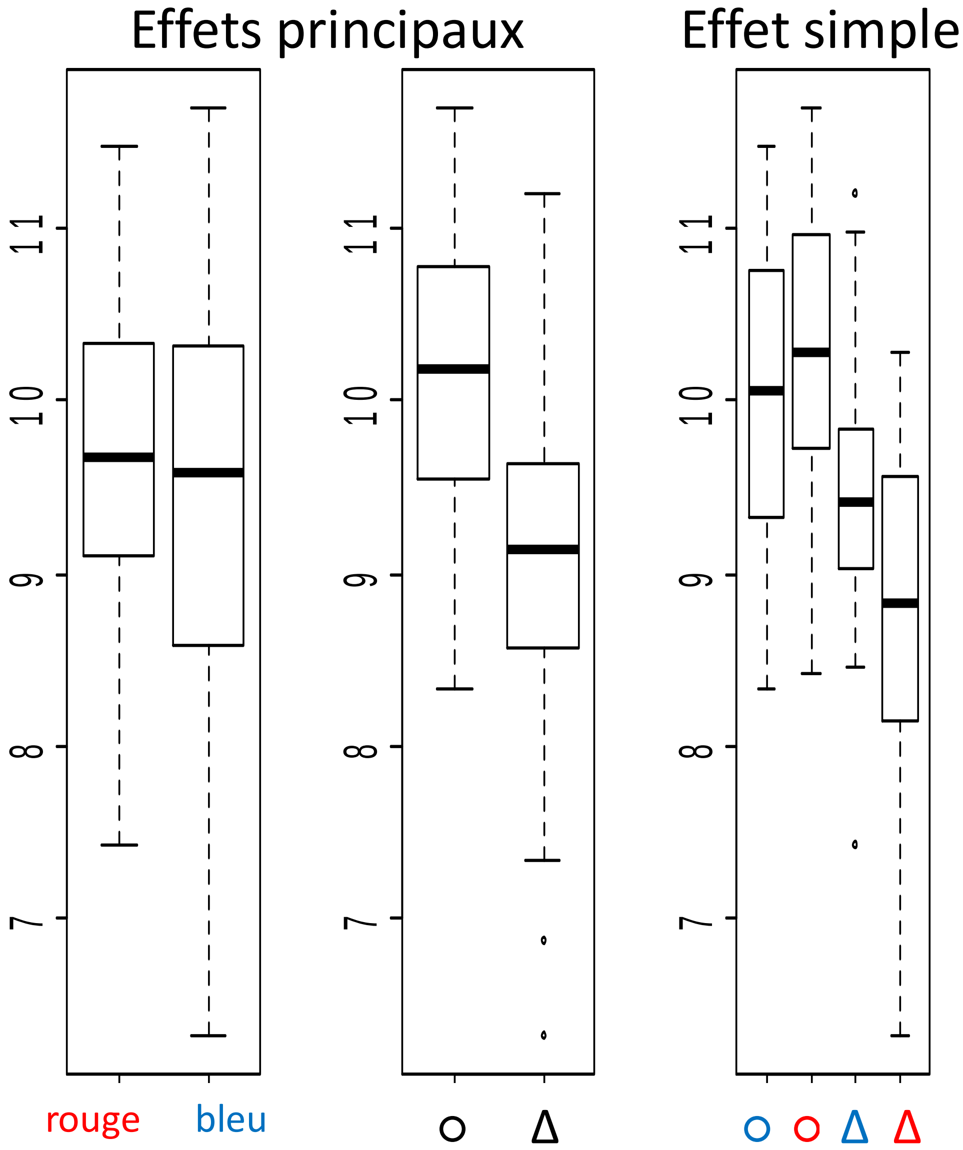
Du point de vue de leur moyenne, ○ < ○ alors que Δ > Δ



Interaction : effet simple et effets principaux

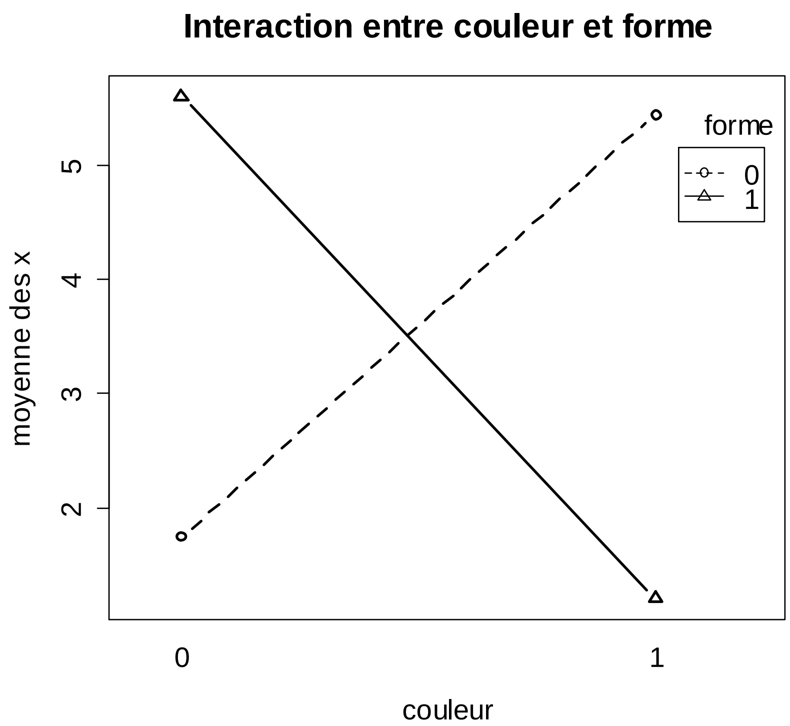


	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
A	1	2.10	2.10	2.537	0.11390	
B	1	32.46	32.46	39.243	6.56e-09	***
A:B	1	6.85	6.85	8.284	0.00476	**
Residuals	116	95.96	0.83			



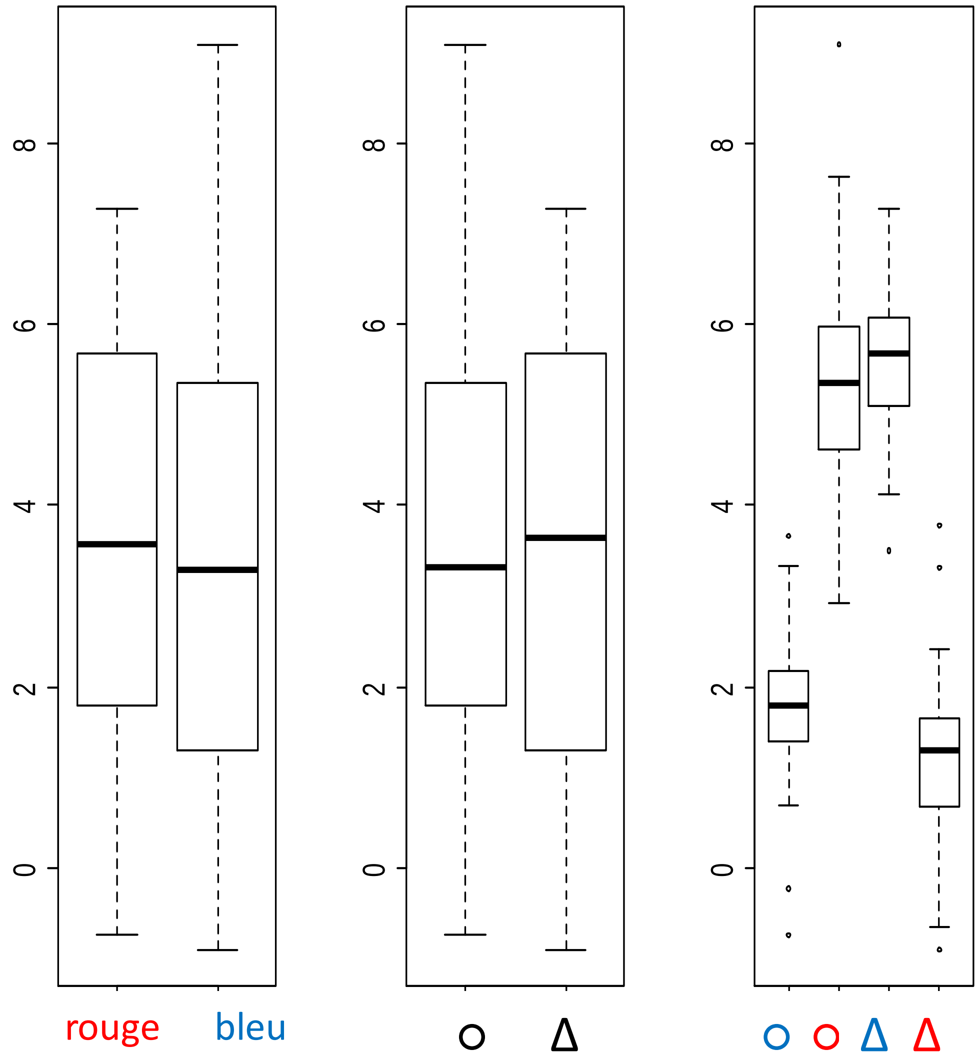
L'absence d'effets principaux peut cacher des effets simples !

```
             Df Sum Sq Mean Sq F value Pr(>F)
A             1    0.0     0.0   0.019  0.892
B             1    0.0     0.0   0.033  0.856
A:B           1  470.0   470.0 505.571 <2e-16 ***
Residuals   116  107.8     0.9
```

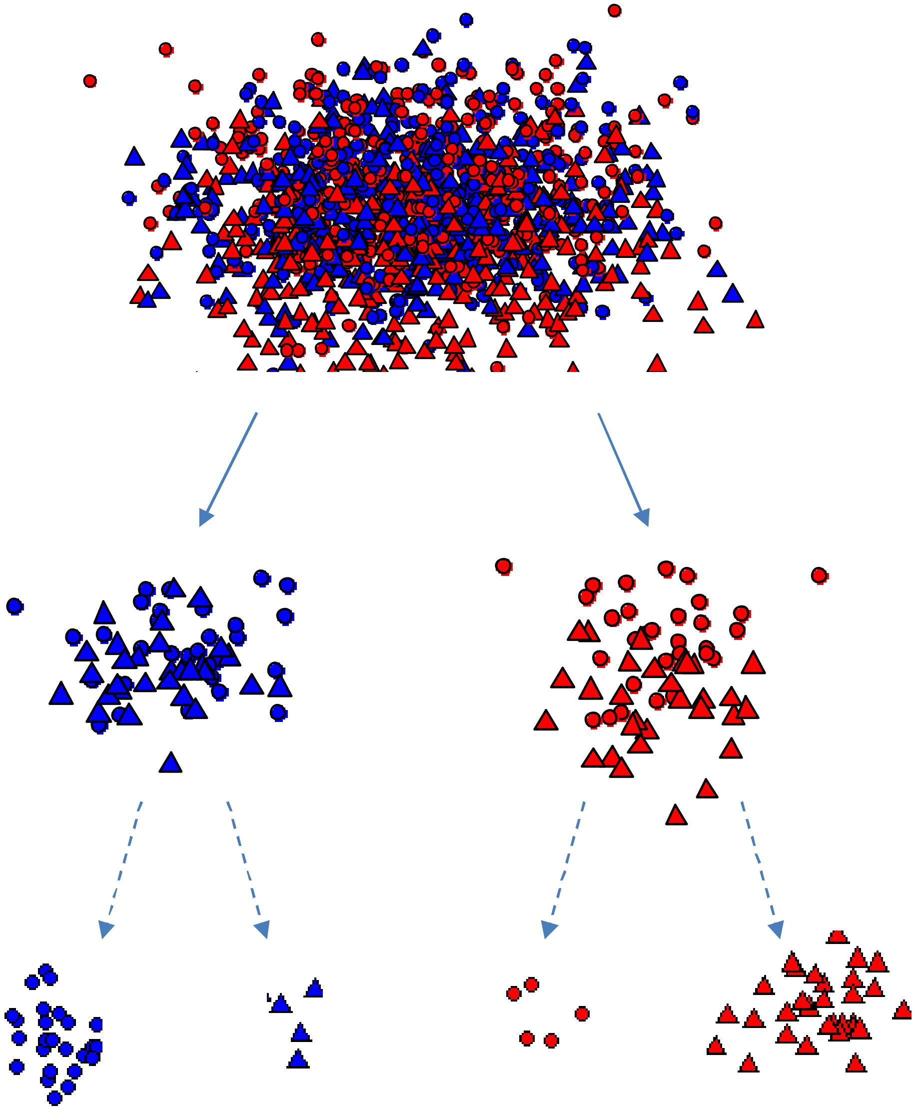


Effets principaux

Effet simple



atoire



Je tire 2 échantillons au hasard,
un bleu et un rouge.

Le facteur couleur est un **facteur fixé** : bleu et rouge sont les deux seules modalités qui m'intéressent et je fixe la taille de mes échantillons.

Le facteur forme est un **facteur aléatoire** : je ne sais pas si les deux formes observées sont les seules dans ma population et je ne peux pas fixer le nombre de ronds ni de triangles.

Beaucoup d’autres concepts attachés à l’ANOVA :

Mesures répétées
Facteurs emboîtés vs croisés
Plan d’expérience
etc.

Are ANOVA and linear regression twin princesses grown in different castles ?

- L'ANOVA c'est pour les variables explicatives qualitatives, la régression linéaire c'est pour les variables explicatives quantitatives.
- L'ANOVA met l'accent sur la comparaison (test d'hypothèse), le modèle linéaire met l'accent sur la prédiction (modèles).
- L'ANOVA est un cas particulier de la régression linéaire généralisée.
- Tout ce qu'on peut faire avec l'ANOVA, on peut le faire avec le modèle linéaire généralisé… peut-être pas tout à fait tout d'après Gelman.
- Mais moi, j'aime bien l'ANOVA.

Merci de votre attention

MeG MARSEILLE