Statistiques sur R

4. Comparaison de plusieurs moyennes
Analyse de variance à un facteur

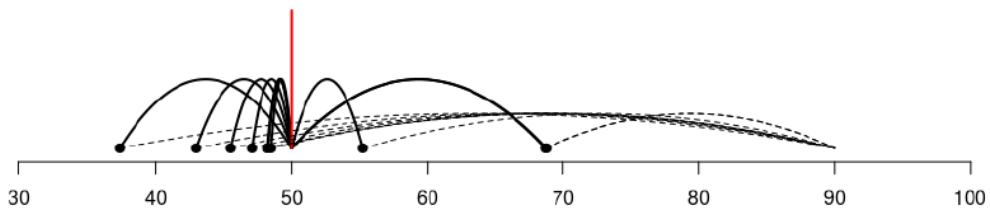
Xavier Bouteiller

bouteiller.xavier@gmail.com

Data University
Institut de science des données

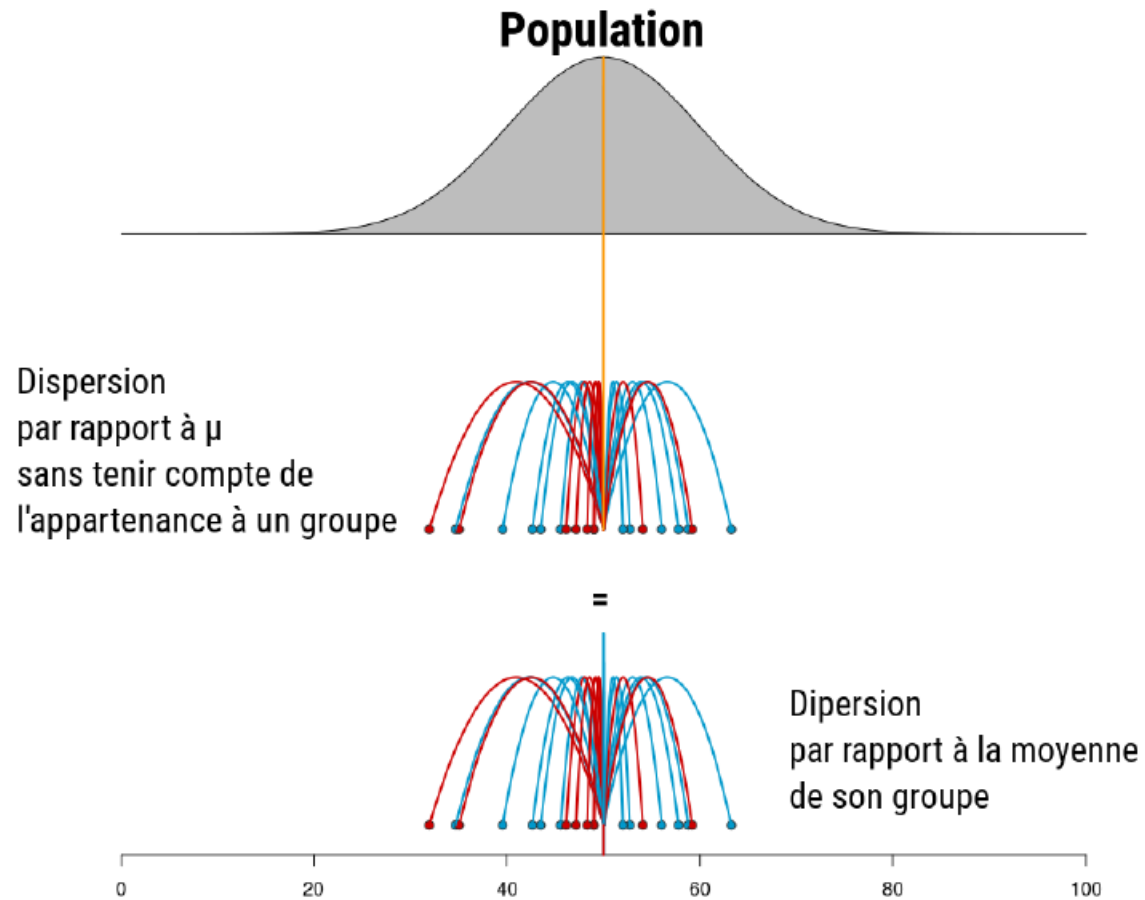
MPDATA

la moyenne est le point de référence qui minimise la variance dans une distribution statistique

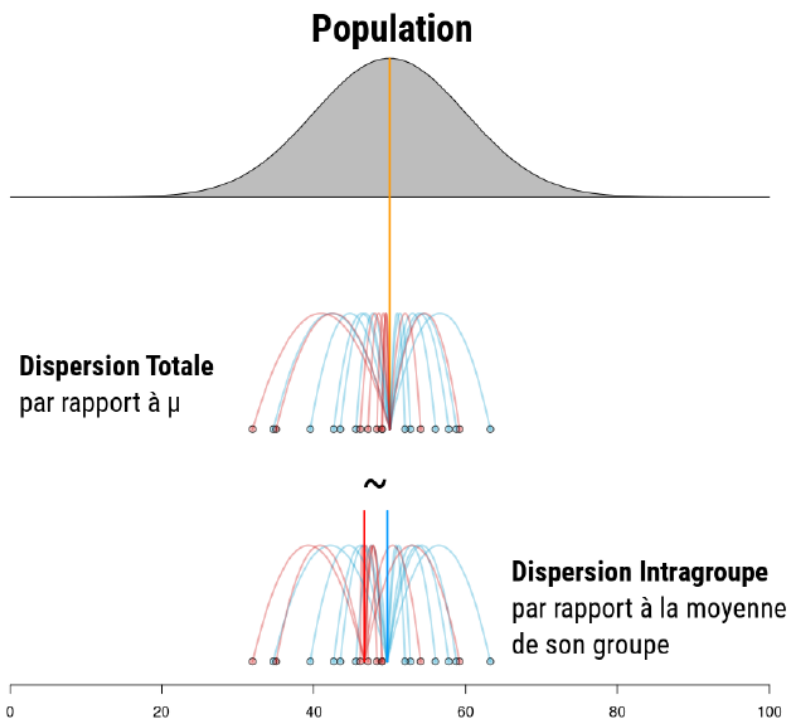


Par rapport à la moyenne (rouge) : 64.13
Par rapport à la valeur 90 (pointillé) : 2077.01

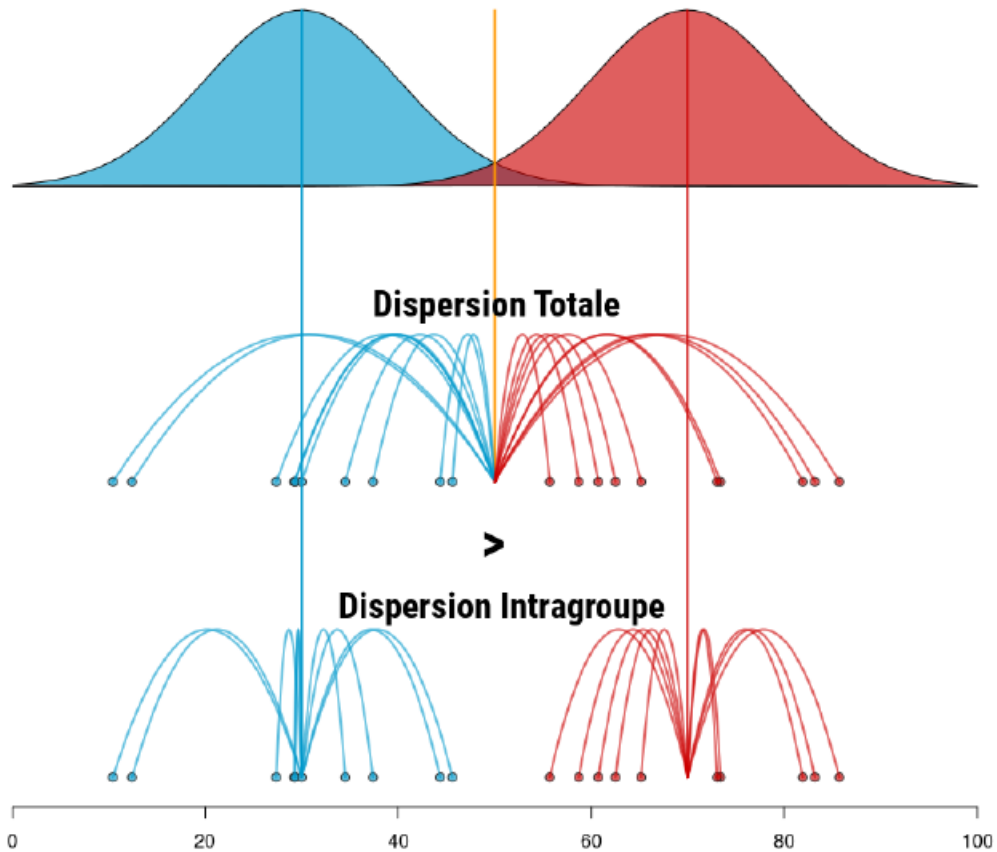
Théoriquement, la dispersion par rapport à μ et par rapport à la moyenne des groupes devraient être égales si $\mu_1 = \mu_2$



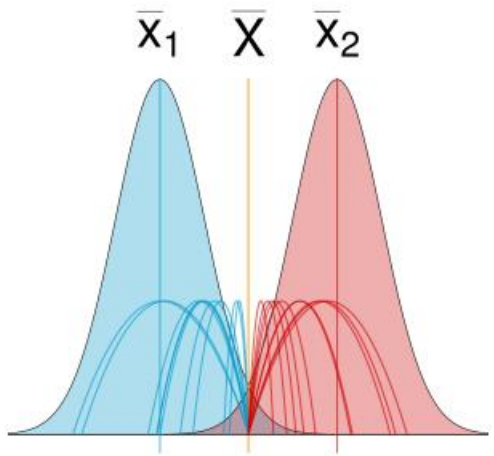
Théoriquement, la dispersion par rapport à μ et par rapport à la moyenne des groupes devraient être égales si $\mu_1 = \mu_2$



Mais, la dispersion par rapport à μ va être très supérieure à la dispersion par rapport à la moyenne des groupes si $\mu_1 \neq \mu_2$



Dispersion Totale
SCT

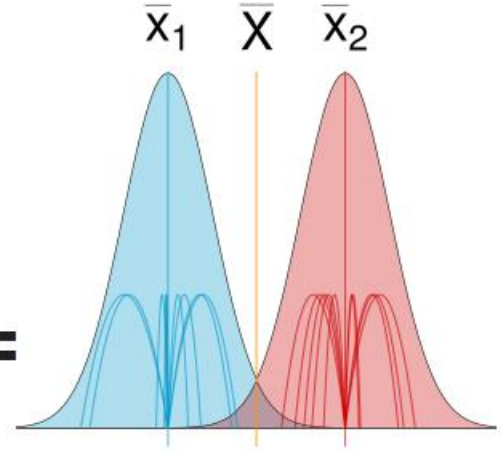


Somme des carrés des écarts à la moyenne générale

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

=

Dispersion Intra-groupe
SCE

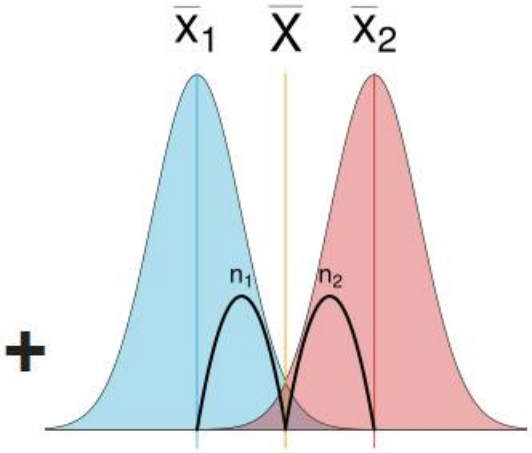


Somme des carrés des écarts à la moyenne du groupe propre

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

+

Dispersion Inter-groupe
SCI



Somme de n_{ij} fois le carré de l'écart entre la moyenne du groupe et la générale

$$\sum_{j=1}^{k} n_j(\bar{x}_j - \bar{X})^2$$

Les variances s'obtiennent en divisant la dispersion par le nombre de degrés de liberté

Sources de variation	Totale	Inter-groupe	Intra-groupe
Dispersions	SCT	SCI	SCE
Nombre de ddl	$n-1$	$k-1$	$n-k$
Variances	$s_x^2 = \frac{SCT}{n-1}$	$V_c = \frac{SCI}{k-1}$	$V_e = \frac{SCE}{n-k}$

Sum of Squares (SS)

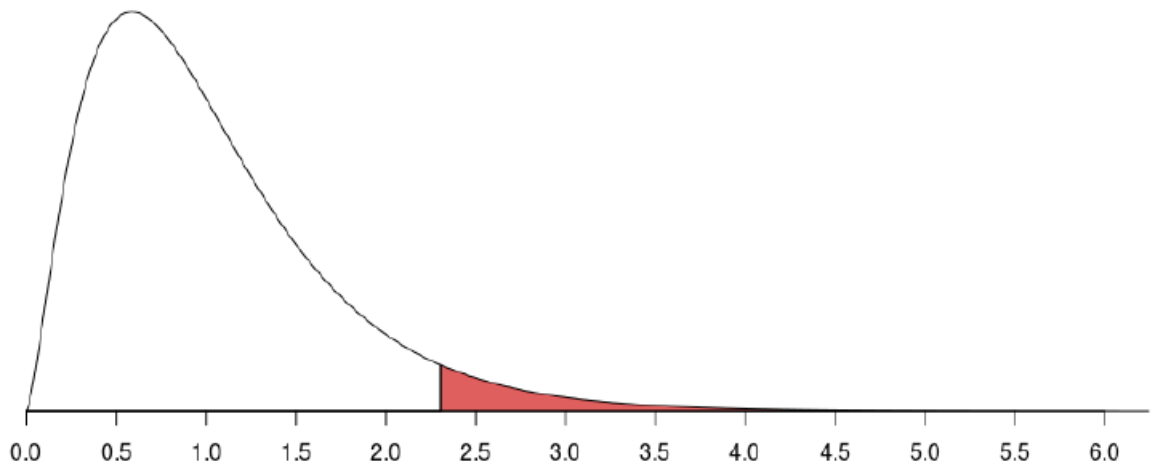
Mean Squares (MS)

Et sous H_0, les variances Inter et Intra sont toutes les 2 égales et sont des estimations de la variance de σ^2
$\rightarrow$ Donc leur rapport est égal à 1

Théoriquement : Si $\mu_1 = \mu_2$ alors $V_c / V_e = 1$

Sous H_0 **:** $F_c = \dfrac{V_c}{V_e}$

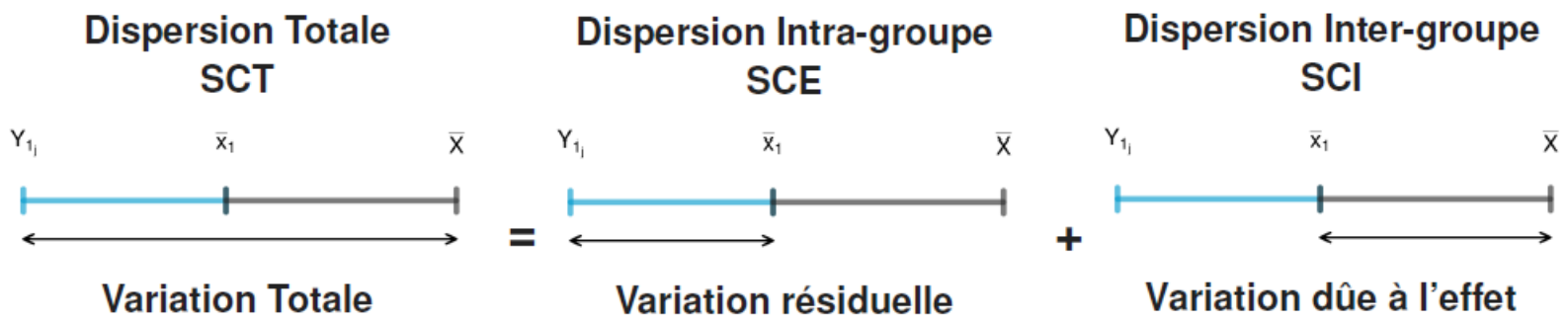
Suit une loi de Fisher à $\upsilon_1 = k - 1$ et $\upsilon_2 = n - k$ degrés de liberté.



L'analyse de la variance :

- s'applique aussi bien aux grands et aux petits échantillons
- vérifie en 1 seul test si les différences observées au niveau des moyennes de k échantillons (2 et plus) sont imputables aux fluctuations d'échantillonnage
- suppose l'égalité des variances
- suppose la normalité des populations d'origine

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_j = \mu_k$
H_1 : Au moins deux moyennes sont différentes.



y_i ε_i α_j

Et donc l'écriture du modèle qui suit :

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim N(0,1)$$

μ : moyenne globale
α_i : correspond à la moyenne de chacun des i groupes
ε_{ij} : erreur résiduelle pour chacun des j individus

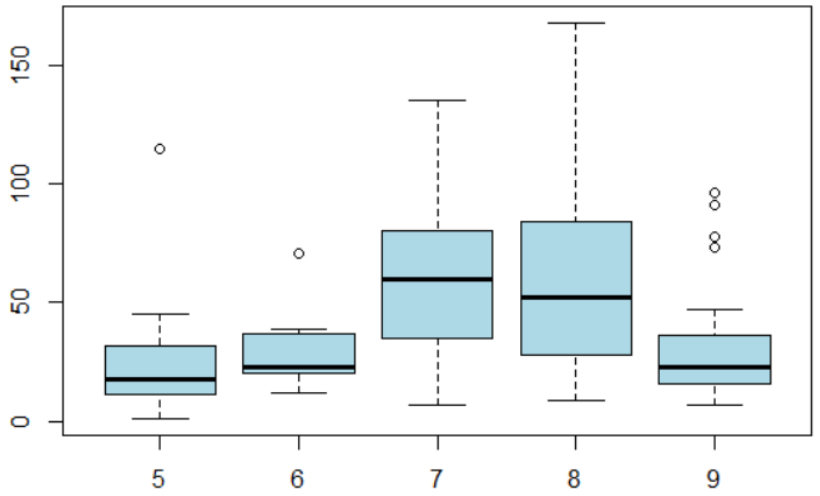
Autrement dit :

H0 : $y_{ij} = \mu + \varepsilon_{ij}$ $\qquad \varepsilon_{ij} \sim N(0,1)$

H1 : $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$

Autrement dit au moins 1 des α_i est différent de 0

Exemple : Ozone dans l'air en fonction des mois



H0 : $y_{ij} = \mu + \varepsilon_{ij}$

H1 : $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$

Sur R :

- μ : intercept
- α_i : les coefficients estimés

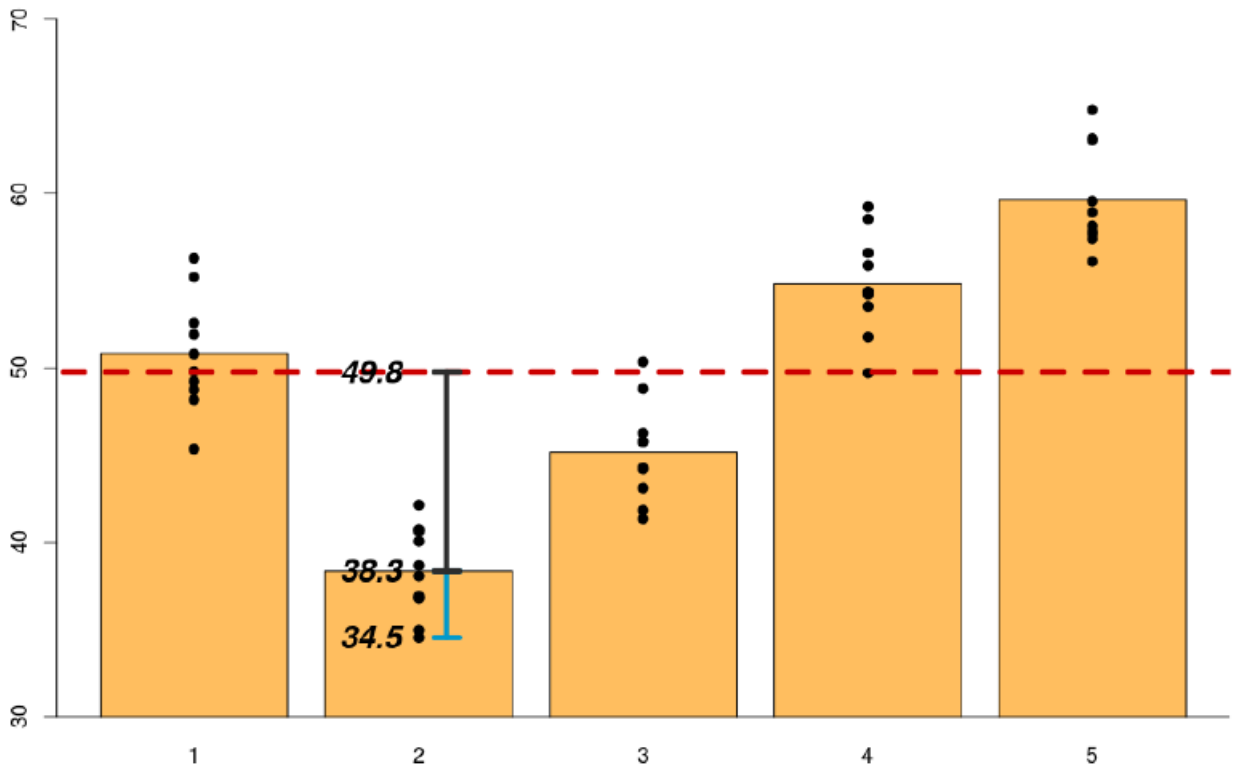
```
> summary(lm(Ozone ~ Month, data = airquality))

Call:
lm(formula = Ozone ~ Month, data = airquality)

Residuals:
    Min      1Q  Median      3Q     Max
-52.115 -16.823  -7.282  13.125 108.038

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   23.615      5.759   4.101 7.87e-05 ***
Month6         5.829     11.356   0.513    0.609
Month7        35.500      8.144   4.359 2.93e-05 ***
Month8        36.346      8.144   4.463 1.95e-05 ***
Month9         7.833      7.931   0.988    0.325
```

Des tailles de lézard dans 5 populations jurasiennes



$$y_{17} = 38.3 - 3.8$$

Dans un modèle d'analyse de variance, comme pour le modèle de la moyenne et le modèle d'analyse de moyenne de Student, **ce sont les résidus qui suivent une loi Normale**.

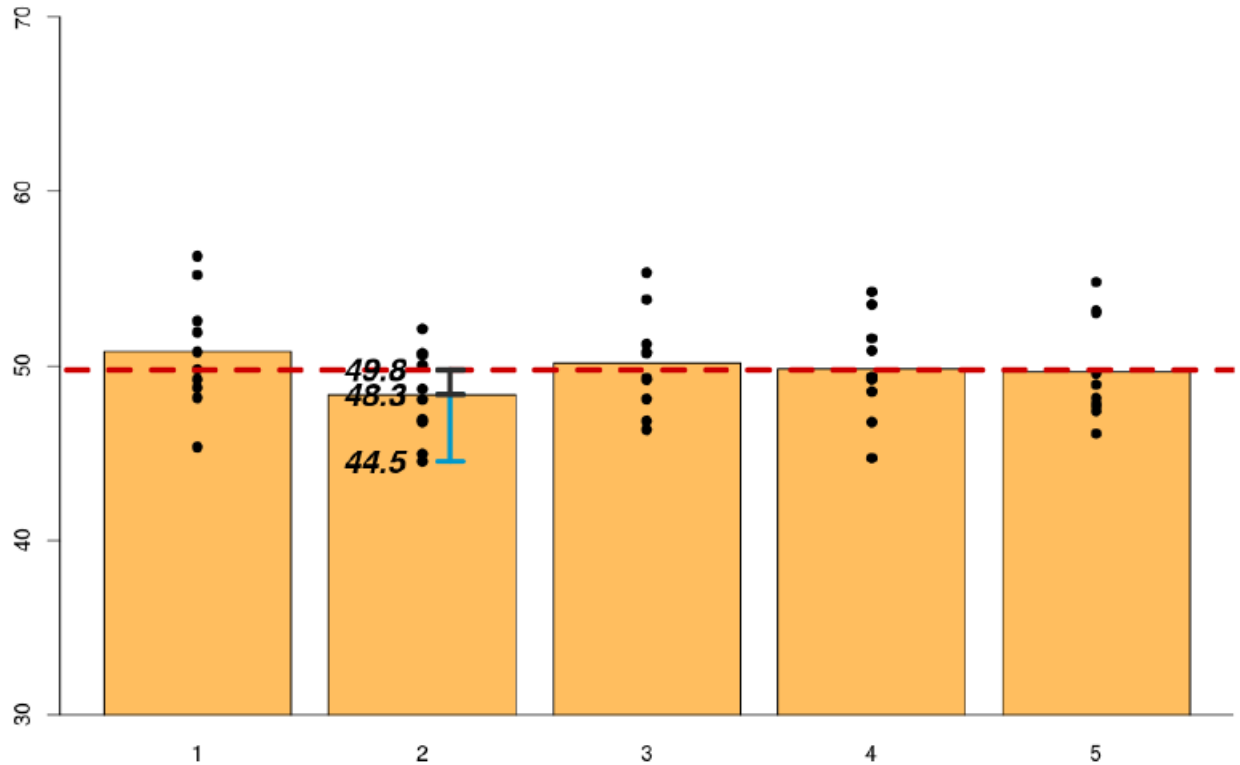
$$y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$$

avec

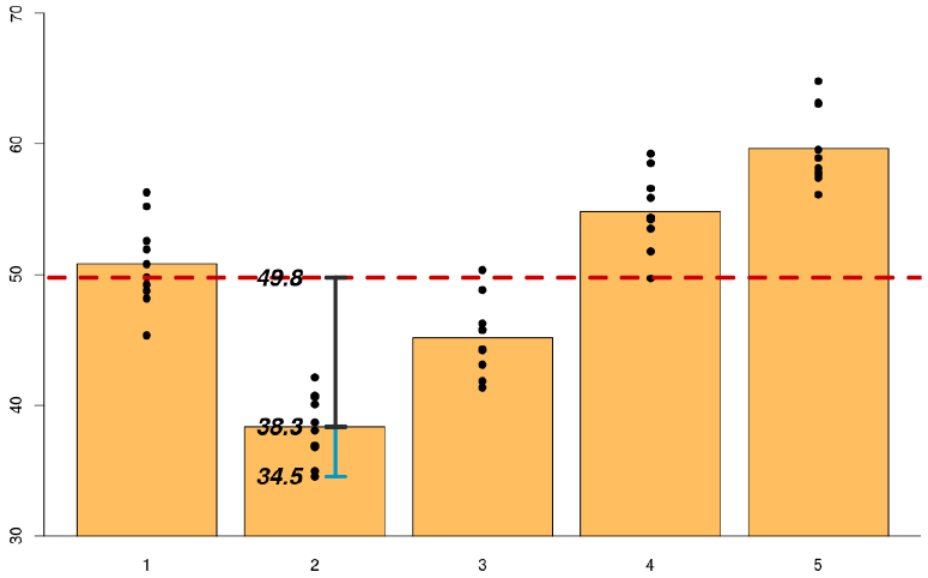
$$\varepsilon_{ij} \sim N(0,1)$$

Donc pour débarasser les variations totales des sources de variations dûes aux différences de moyenne :

$$\varepsilon_{ij} = y_{ij} - \mu - \alpha_j$$



$$y_{17} = 48.3 - 3.8$$

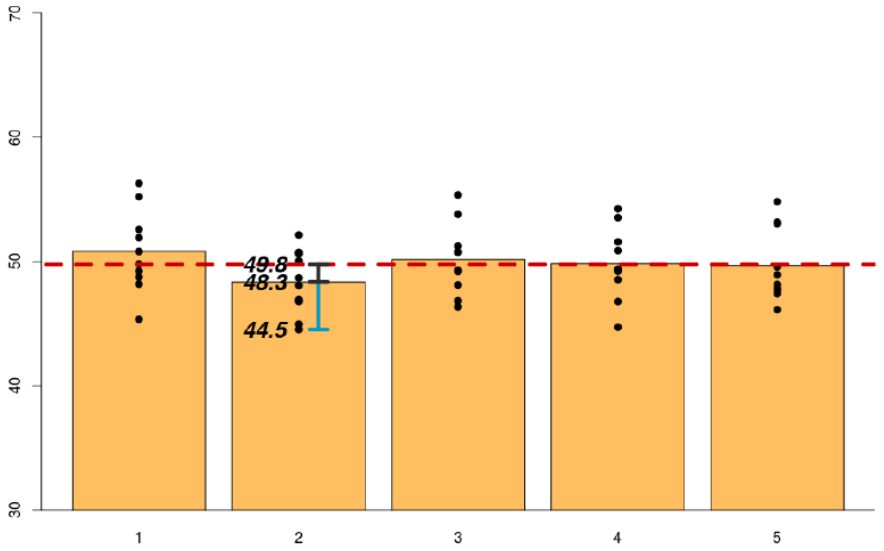


Sources de variation	Totale	Inter-groupe	Intra-groupe
Dispersions	3147	2761.7	385.3
Nombre de ddl	$n-1$	$k-1$	$n-k$
Variances	$s_x^2 =$	$V_c =$	$V_e =$

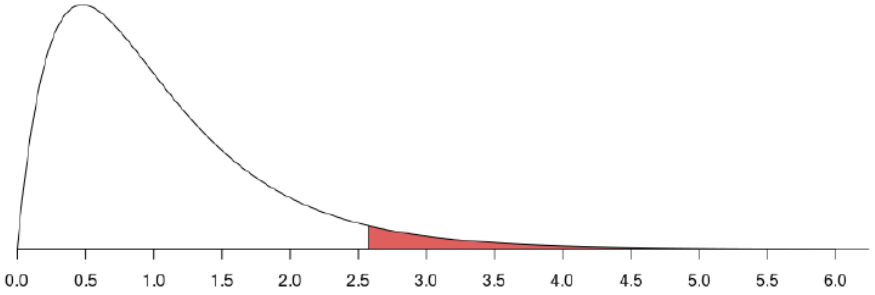
Théoriquement : Si $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ alors $V_c / V_e = 1$

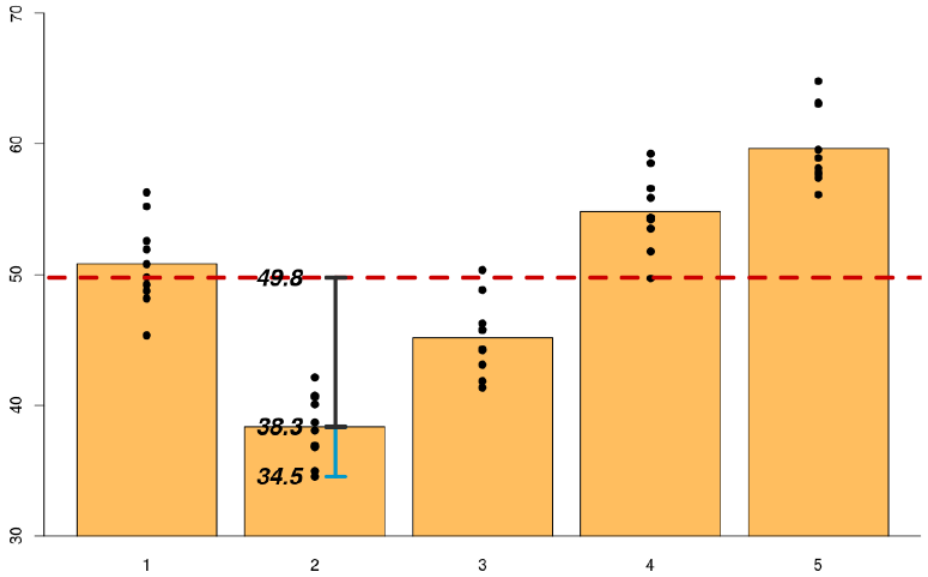
Sous H_0 : $F_c = \dfrac{V_c}{V_e}$

Suit une loi de Fisher à $\upsilon_1 = k-1$ et $\upsilon_2 = n-k$ degrés de liberté.



Sources de variation	Totale	Inter-groupe	Intra-groupe
Dispersions	418.1	32.8	385.3
Nombre de ddl	$n-1$	$k-1$	$n-k$
Variances	$s_x^2 =$	$V_c =$	$V_e =$



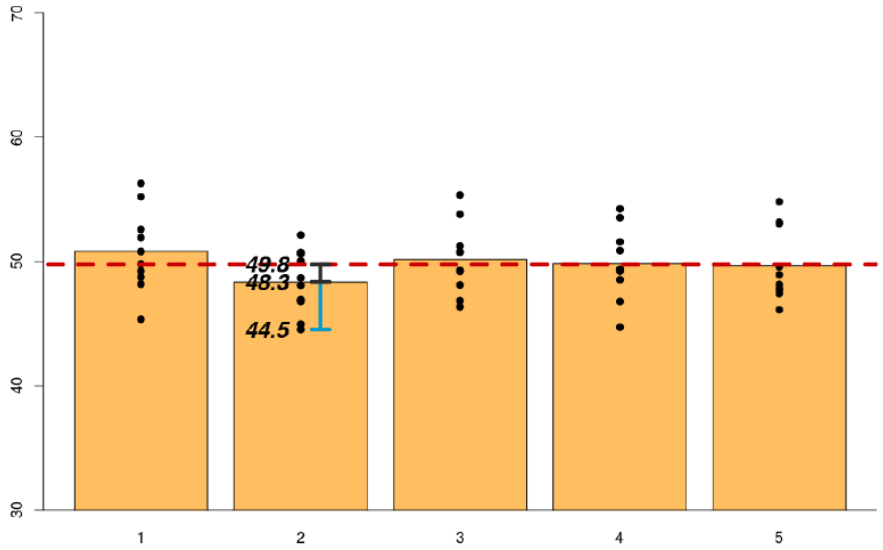


Sources de variation	Totale	Inter-groupe	Intra-groupe
Dispersions	3147	2761.7	385.3
Nombre de ddl	$n-1$	$k-1$	$n-k$
Variances	$s_x^2 =$	$V_c =$	$V_e =$

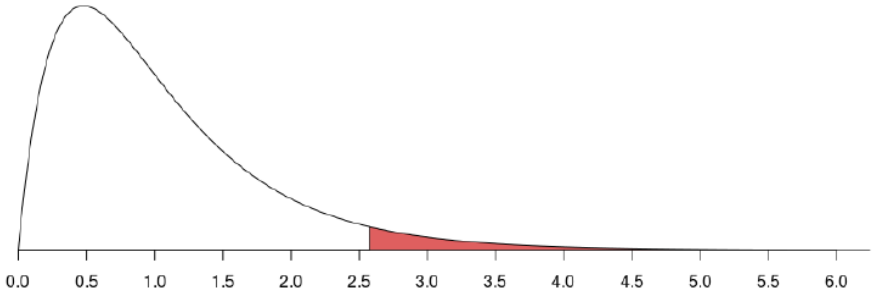
Théoriquement : Si $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ alors $V_c / V_e = 1$

Sous H_0 : $F_c = \dfrac{V_c}{V_e}$

Suit une loi de Fisher à $\upsilon_1 = k-1$ et $\upsilon_2 = n-k$ degrés de liberté.



Sources de variation	Totale	Inter-groupe	Intra-groupe
Dispersions	418.1	32.8	385.3
Nombre de ddl	$n-1$	$k-1$	$n-k$
Variances	$s_x^2 =$	$V_c =$	$V_e =$



Si H_0 est rejetée, il s'agira ensuite de déterminer quels sont les couples de moyennes d'échantillons qui sont différentes.
$\rightarrow$ Test post-hoc vu en TDM.

Let’s practice …