Modelos Lineares

Distribuições de Probabilidades
Distribuição de Student, Chi2, Fischer
Testes de Aderência à Lei Normal
Análise de Variância

Professora Ariane Ferreira

Instituto Politécnico
Campus Regional da UERJ
Nova Friburgo - RJ

(t) Student Distribution

- Useful sampling distribution based on the normal distribution.
- If X and χ^2_k are standard normal and chi-square random variables, then t_k is distributed with k degrees freedom.

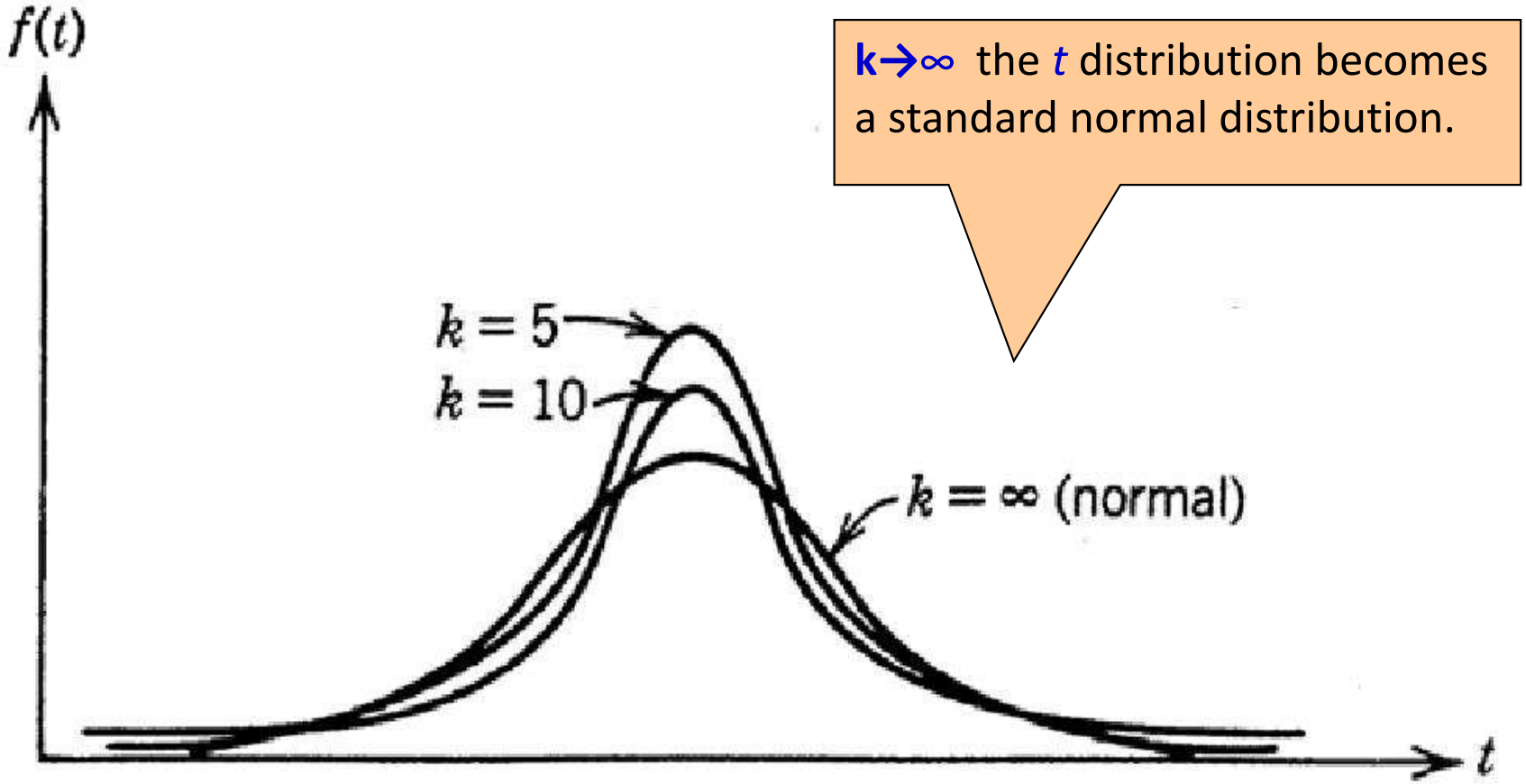
$$t_k \cong \frac{x}{\sqrt{\chi^2_k / k}}$$

(t) Student Probability density function

$$f(t)=\frac{\Gamma\left[\left(K+1\right)/2\right]}{\sqrt{K\pi\left(k/2\right)}}\left(\frac{t^2}{k}+1\right)^{-\left(k+1\right)/2}$$

- For a random sample of size *n* collected from a $N(\mu,\sigma^2)$ distribution with a sample mean and a sample variance, $\left(\bar{x},s^2\right)$
- It can be shown that: $\dfrac{\bar{x}-\mu}{s/\sqrt{n}}\cong t_{n-1}$
- The *t* distribution is used to make inferences about the mean of normal distribution.

Several (t) student Distributions



k→∞ the *t* distribution becomes a standard normal distribution.

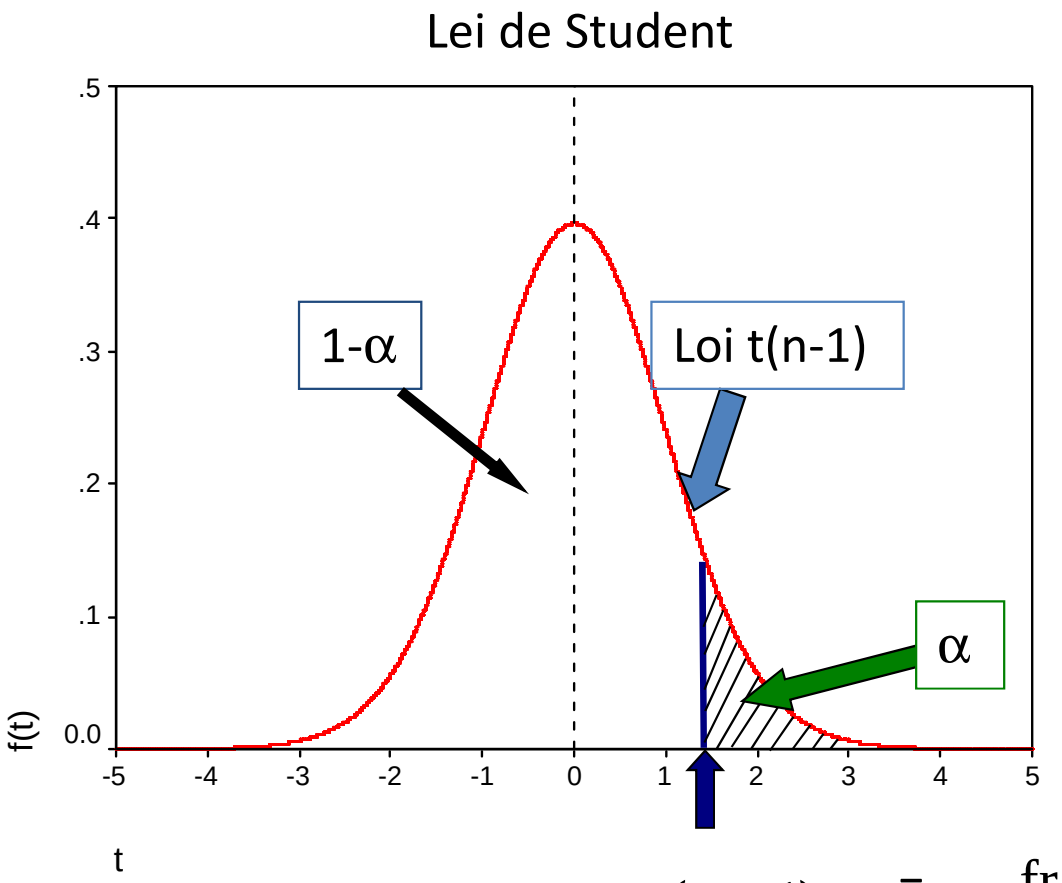
Lei de Student

Se $X \sim N(\mu, \sigma)$ então :

$$T = \frac{\bar{X} - \mu}{S / \sqrt{n}}$$

T segue uma distribuição de Student com *n-1* graus de liberdade [notação $t(n\text{-}1)$].

Fractis da Lei de Student



$$t(n-1) \Rightarrow N(0,1)$$

lorsque $n \Rightarrow \infty$

$t_{1-\alpha}(n-1)$ = fractile d'ordre $1-\alpha$ d'une loi de Student à n-1 degrés de liberté

Fractiles de la loi de Student

ν \ P	0,60	0,70	0,80	0,90	0,95	0,975	0,990	0,995	0,999	0,9995
1	0,325	0,727	1,376	3,078	6,314	12,71	31,82	63,66	318,3	636,6
2	0,289	0,617	1,061	1,886	2,920	4,303	6,965	9,925	22,33	31,60
3	0,277	0,584	0,978	1,638	2,353	3,182	4,541	5,841	10,22	12,94
4	0,271	0,569	0,941	1,533	2,132	2,776	3,747	4,604	7,173	8,610
5	0,267	0,559	0,920	1,476	2,015	2,571	3,365	4,032	5,893	6,859
6	0,265	0,553	0,906	1,440	1,943	2,447	3,143	3,707	5,208	5,959
7	0,263	0,549	0,896	1,415	1,895	2,365	2,998	3,499	4,785	5,405
8	0,262	0,546	0,889	1,397	1,860	2,306	2,896	3,355	4,501	5,041
9	0,261	0,543	0,883	1,383	1,833	2,262	2,821	3,250	4,297	4,781
10	0,260	0,542	0,879	1,372	1,812	2,228	2,764	3,169	4,144	4,587
11	0,260	0,540	0,876	1,363	1,796	2,201	2,718	3,106	4,025	4,437
12	0,259	0,539	0,873	1,356	1,782	2,179	2,681	3,055	3,930	4,318
13	0,259	0,538	0,870	1,350	1,771	2,160	2,650	3,012	3,852	4,221
14	0,258	0,537	0,868	1,345	1,761	2,145	2,624	2,977	3,787	4,140
15	0,258	0,536	0,866	1,341	1,753	2,131	2,602	2,947	3,733	4,073
16	0,258	0,535	0,865	1,337	1,746	2,120	2,583	2,921	3,686	4,015
17	0,257	0,534	0,863	1,333	1,740	2,110	2,567	2,898	3,646	3,965
18	0,257	0,534	0,862	1,330	1,734	2,101	2,552	2,878	3,611	3,922
19	0,257	0,533	0,861	1,328	1,729	2,093	2,539	2,861	3,579	3,883
20	0,257	0,533	0,860	1,325	1,725	2,086	2,528	2,845	3,552	3,850
21	0,257	0,532	0,859	1,323	1,721	2,080	2,518	2,831	3,527	3,819
22	0,256	0,532	0,858	1,321	1,717	2,074	2,508	2,819	3,505	3,792
23	0,256	0,532	0,858	1,319	1,714	2,069	2,500	2,807	3,485	3,767
24	0,256	0,531	0,857	1,318	1,711	2,064	2,492	2,797	3,467	3,745
25	0,256	0,531	0,856	1,316	1,708	2,060	2,485	2,787	3,450	3,725
26	0,256	0,531	0,856	1,315	1,706	2,056	2,479	2,779	3,435	3,707
27	0,256	0,531	0,855	1,314	1,703	2,052	2,473	2,771	3,421	3,690
28	0,256	0,530	0,855	1,313	1,701	2,048	2,467	2,763	3,408	3,674
29	0,256	0,530	0,854	1,311	1,699	2,045	2,462	2,756	3,396	3,659
30	0,256	0,530	0,854	1,310	1,697	2,042	2,457	2,750	3,385	3,646
32	0,256	0,530	0,853	1,309	1,694	2,037	2,449	2,738	3,365	3,622
34	0,255	0,529	0,852	1,307	1,691	2,032	2,441	2,728	3,348	3,601
36	0,255	0,529	0,852	1,306	1,688	2,028	2,434	2,719	3,333	3,582
38	0,255	0,529	0,851	1,304	1,686	2,024	2,429	2,712	3,319	3,566
40	0,255	0,529	0,851	1,303	1,684	2,021	2,423	2,704	3,307	3,551
50	0,255	0,528	0,849	1,298	1,676	2,009	2,403	2,678	3,261	3,496
60	0,254	0,527	0,848	1,296	1,671	2,000	2,390	2,660	3,232	3,460
70	0,254	0,527	0,847	1,294	1,667	1,994	2,381	2,648	3,211	3,435
80	0,254	0,527	0,846	1,292	1,664	1,990	2,374	2,639	3,195	3,415
90	0,254	0,526	0,846	1,291	1,662	1,987	2,368	2,632	3,183	3,402
100	0,254	0,526	0,845	1,290	1,660	1,984	2,365	2,626	3,174	3,389
200	0,254	0,525	0,843	1,286	1,653	1,972	2,345	2,601	3,131	3,339
500	0,253	0,525	0,842	1,283	1,648	1,965	2,334	2,586	3,106	3,310
∞	0,253	0,524	0,842	1,282	1,645	1,960	2,326	2,576	3,090	3,291

(χ^2) Chi-Square (Pearson) Distribution

- Important sampling distribution which originates from normal distribution.
- If $X_1, X_2, \ldots, X_\nu$ are ν independent normally distributed random variables with m = 0 and $\sigma^2 = 1$, then the random variable:

$$\chi_\nu^2 = X_1^2 + X_2^2 + \cdots + X_\nu^2$$

- Is distributed as chi-square with ν degrees of freedom.

(χ^2) Chi-Square probability density function

Γ is the gamma function.

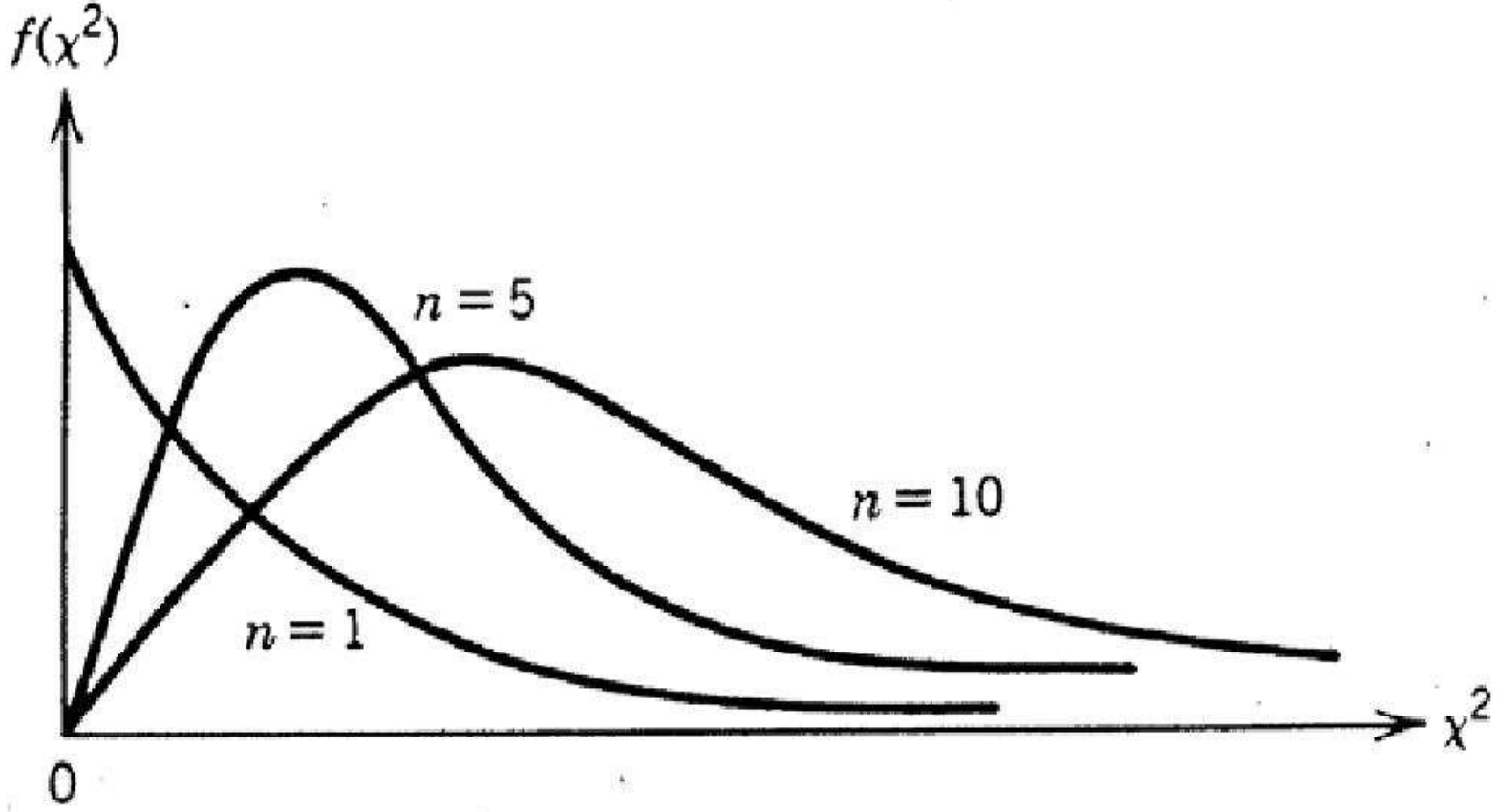
$$f(\chi^2) = \frac{1}{2^{n/2}\Gamma\left(\frac{n}{2}\right)}(\chi^2)^{(n/2)-1}e^{\left(-\chi^2/2\right)}$$

- If a random sample of size n take from a $N(\mu,\sigma^2)$ distribution, and this sample yields a sample variance of s^2 , it can be show that:

χ^2 is used to make inferences about the variance of a normal distribution.

$$\frac{(n-1)s^2}{\sigma^2} \approx \chi^2_{n-1}$$

Several (χ^2) Chi-Square Distribution



(Fisher-Snedecor) F-Distribution

- The F-distribution to be considered based on chi-square distribution.
- If χ^2_u and χ^2_v are chi-square random variables with u and v degrees of freedom, then the ratio:

$$F_{u,v} \equiv \frac{\chi^2_u / u}{\chi^2_v / v}$$

- Is distributed as F with u and v degrees of freedom.

(F) Probability density function

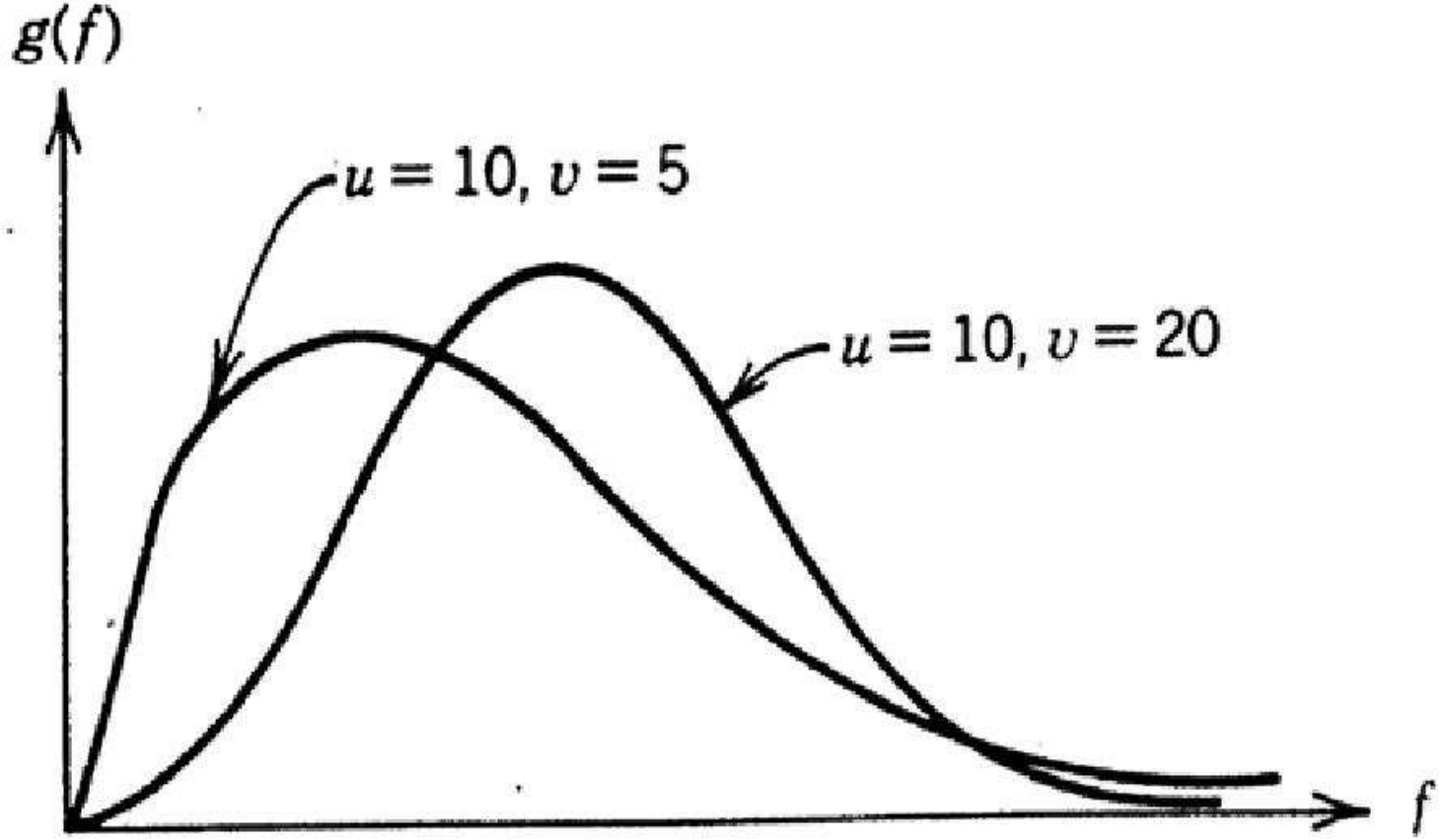
$$g(F)=\frac{\Gamma\left(\frac{u+v}{2}\right)\left(\frac{u}{v}\right)^{\frac{u}{2}}}{\Gamma\left(\frac{u}{2}\right)\Gamma\left(\frac{v}{2}\right)}\frac{F^{\left(u/2-1\right)}}{\left[\left(\frac{u}{2}\right)F+1\right]^{\left(u+v\right)/2}}$$

- For two independent normal processes $x_1 \sim N(\mu_1, \sigma_1^2)$ and $x_2 \sim N(\mu_2, \sigma_2^2)$ with two random samples of sizes n_1 and n_2 , yield sample variances s_1^2 and s_2^2 , it can be shown that:

$$\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \sim F_{n_1-1,n_2-1}$$

The *F* distribution is used to make inferences in comparing the variances of two normal distribution.

Several (F) Distributions



Testes d’hypothèses statistiques

- Quand on veut démontrer l’hypothèse H1 que :
  - une moyenne mesurée dans un échantillon est significativement différente de la moyenne dans la population
  - significativement = ne résulte pas uniquement du hasard
  - des moyennes mesurées dans 2 échantillons sont significativement différentes
  - une variable ne suit pas une loi théorique donnée
  - deux variables sont significativement différentes
  - un échantillon n’est pas homogène mais est composé de plusieurs sous-populations

Principaux Testes d'hypothèses statistiques

- Égalité de moyennes dans 2 échantillons : test de Student
- Égalité de moyennes dans k > 2 échantillons : analyse de la variance
- Égalité de 2 variances : test de Fisher-Snedecor
- Égalité de 2 distributions : test de Kolmogorov-Smirnov
- Indépendance de 2 variables qualitatives : test du χ^2
  - ce test est non-paramétrique
  - mais non exact -> le test exact correspondant est le test de Fisher (ne pas confondre avec le test de Fisher-Snedecor)
- voir plus loin ces notions de « paramétrique » et « exact »

Comment réaliser un test d'hypothèse

Soit $\{x_1,\ldots,x_n\}$ un échantillon de ***n*** réalisations indépendantes de la v.a. ***X***.
Soit **F(x)** la loi de distribution inconnue de ***X***.
L'hypothèse de départ sera que la loi de distribution est $\mathbf{F_X(x).}$

$$H_0 : F_X(x) = F(x)$$
$$H_1 : F_X(x) \neq F(x)$$

- on soumet l'hypothèse contraire H_0 à un test T qui doit être satisfait si H_0 est vraie
- puis on montre que T n'est pas satisfait ⇒ H_0 est faux
- Vocabulaire : H_0 : hypothèse nulle – H_1 : hypothèse alternative
- À l'hypothèse nulle H_0 est associée une statistique, fonction des observations, qui suit une loi théorique connue si H_0 est vraie

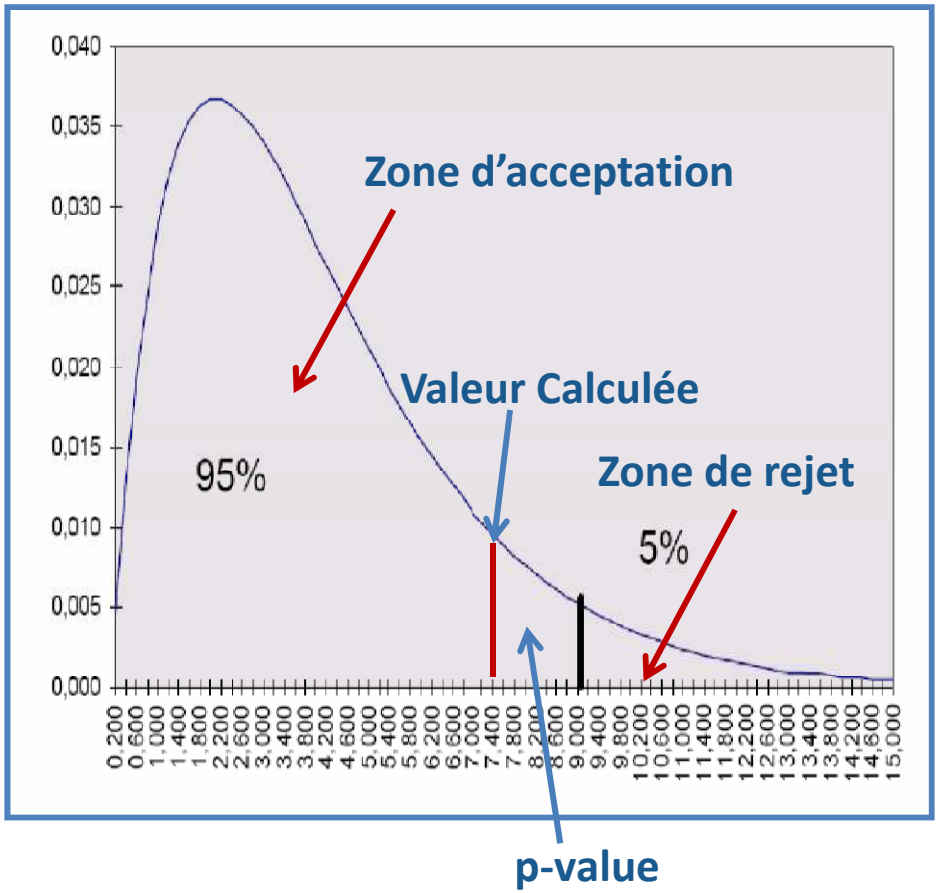
Comment réaliser un test d'hypothèse

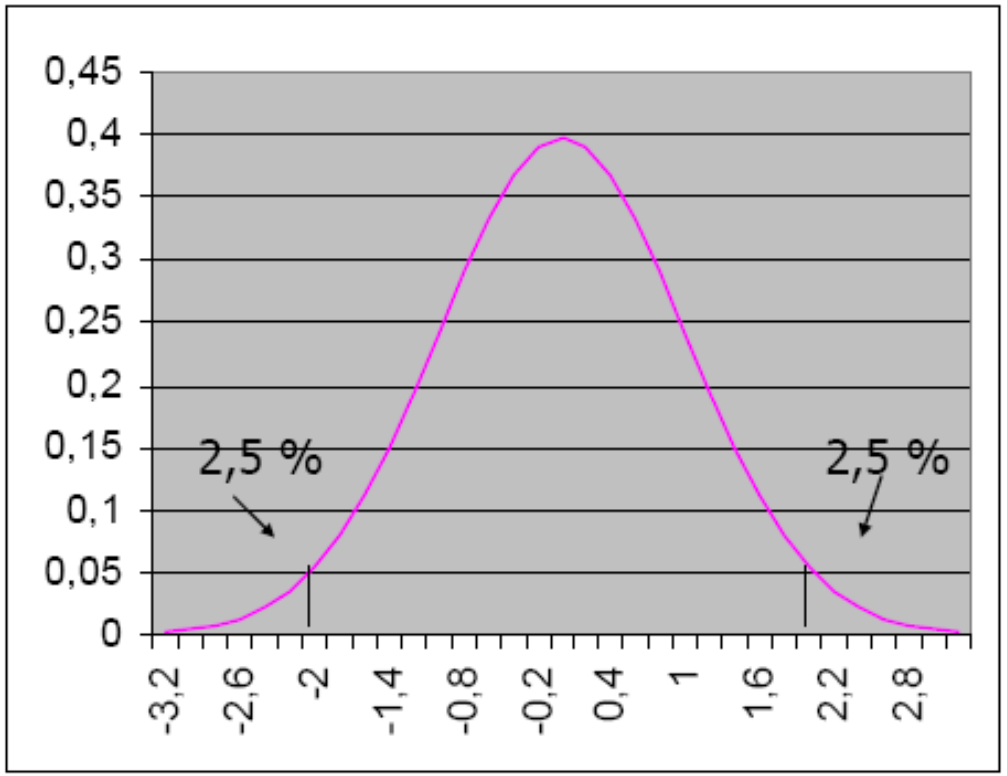
Exemple : si l'hypothèse nulle est ($H_0 : \mu = \mu_0$), alors $\frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$ suit une loi normale réduite (n grand)

Test d'hypothèse

$$H_0 : \mu = \mu_0$$

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





Comment réaliser un test d'hypothèse

- **Loi de distribution du test**
  - Par rapport à la loi de distribution du test, choisir une zone de rejet (unilatérale ou bilatérale),
- **Zone de Rejet**
  - Le test est caractérisée par une probabilité α d'être dans cette zone :
  - on choisit souvent **α = 0,05 (= 5 %)**
- **Zone de d'acceptation**
  - le complémentaire est la zone d'acceptation (si α = 0,05, il s'agit de la région autour de la moyenne où se trouvent 95 % des valeurs de la statistique).
- **Quoi faire?**

- Mesurer la valeur de la statistique sur l'échantillon et comparer cette valeur aux valeurs théoriques de la loi de distribution.
  - Si cette valeur mesurée tombe dans la zone de rejet, on rejette H_0
  - Sinon, on ne la rejette pas

Définitions importantes dans un test d'hypothèse

- **Niveau de signification**
  - Il est le degré de signification du test;
  - Représenté par le p-value.
- **p-value**
  - probabilité d'obtenir une statistique de test aussi extrême (≥ ou ≤) que la valeur mesurée sur l'échantillon si H_0 est vraie.
  - p-value ≥ α ⇒ ne pas rejeter H_0
  - p-value < α ⇒ rejeter H_0 (on considère qu'il est trop peu probable d'avoir une si faible p-value si H_0 est vraie, pour admettre que H_0 est vraie)
- **Avantage de la p-value**
  - elle a un sens absolu, qui ne dépend pas de la loi de probabilité et du nombre de degrés de liberté.

Types d'erreurs d'un test d'hypothèses

- **Deux erreurs possibles :**
  - le **rejet** d'une **H_0 vraie** (risque de 1ère espèce, ou de **Type I**)
    - probabilité de cette erreur = **α**
  - le **non rejet** d'une **H_0 fausse** (risque de 2de espèce, ou de **Type II**)
    - probabilité de cette erreur = **β**

Les différentes situations que l'on peut rencontrer dans le cadre des tests d'hypothèse sont résumées dans le tableau suivant :

Décision \ Réalité	H_0 vraie	H_0 fausse
Non-rejet de H_0	correct **(1-α)**	Manque de puissance **risque de second espèce β**
Rejet de H_0	Rejet à tort **risque de première espèce α**	Puissance du test **(1 - β)**

Rappel : Testes d'hypothèses statistiques

- **On ne peut réduire simultanément α et β**
- **Puissance d'un test** : 1 – risque β
  - Probabilité de rejeter H_0 si celle-ci est fausse ⇒ décision correcte.
- **Le risque β et la puissance 1 – β dépendent de :**
  - la vraie valeur du paramètre de la population (plus elle est éloignée de la valeur testée, plus le risque β baisse)
  - l'écart-type σ de la population (σ ⇑ ⇒ β ⇓)
  - le risque α choisi (α ⇑ ⇒ β ⇓)
  - la taille n de l'échantillon (n ⇑ ⇒ β ⇓)
- **La puissance d'un test augmente avec la taille de l'échantillon**
  - plus les observations sont nombreuses, plus on a d'éléments permettant de rejeter H_0 si elle est fausse.
- **Attention :**
  - avec des tests puissants, on rejette facilement H_0 dès que le nombre d'observations augmente :
    - exemples : le test du χ^2, les tests de normalité

Remarques sur les tests d'hypothèses

- **Remarques :**
  - les tests d'hypothèse s'appliquent bien à des hypothèses H_0 contraignantes :

Test d'hypothèse

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

  - car elles conduisent à des tests T précis.
  - les tests permettent de prouver qu'un échantillon est hétérogène ou n'a pas été constitué par un tirage au hasard, mais non l'inverse.

Tests d'hypothèses paramétriques et non-paramétriques

- **Tests paramétriques :**
  - supposent que les variables suivent une loi particulière (normalité, homoscédasticité)
  - parfois plus puissants que des tests non-paramétriques, mais rarement beaucoup plus
  - exemples : test de Student, ANOVA
- **Tests non-paramétriques :**
  - ne supposent pas que les variables suivent une loi particulière;
  - se fondent souvent sur les rangs des valeurs des variables plutôt que sur les valeurs elles-mêmes;
  - Ils sont peu sensibles aux valeurs aberrantes;
  - à privilégier avec de petits effectifs (< 10);
  - par définition, les tests d'adéquation à une loi (ex : tests de normalité) sont non-paramétriques.

Tests d'hypothèses exacts et asymptotiques

- **Test exact :**
  - utilisable sur des données éparses
  - calcul direct de probabilité
  - prennent en compte tous les cas de figure possibles
  - calcul pouvant être coûteux en temps machine
    - variante : approximation par la méthode de Monte-Carlo
    - exemple : test de Fisher
- **Tests asymptotique :**
  - approximation valable quand les effectifs sont assez grands et les tableaux de données assez denses
  - exemple : test du χ^2 (si effectifs théoriques ≥ 5)

Tests d'adéquation à une loi normale $X \sim N(\mu, \sigma^2)$

Unidimensionnelle

- **Anderson-Darling**
- **Kolmogorov-Smirnov**
- **Normal Q-plot**

Multidimensionnelle

- **Mardia**
- **Multivarié normal Q-Plot**

Test d'hypothèse

Soit $\{x_1,\ldots,x_n\}$ un échantillon de n réalisations indépendantes de la v.a. X.
Soit $F(x)$ la loi de distribution inconnue de X.
L'hypothèse de départ sera que la loi de distribution est $F_X(x)$.

$$H_0 : F_X(x) = F(x)$$
$$H_1 : F_X(x) \neq F(x)$$

Test d'Anderson-Darling

- Calculer $\bar{x}$ et s
- **Calculer la probabilité :**

$$z_{(k)} = \Phi\left(\frac{x_{(k)} - \bar{x}}{s}\right)$$

- **Calculer la statistique** :

$$A = -\frac{1}{n}\left\{\sum_{k=1}^{n}(2k-1)\left\{\ln z_{(k)} + \ln(1 - z_{(n+1-k)})\right\}\right\} - n$$

$$A^* = \left(1 + \frac{0.75}{n} + \frac{2.25}{n^2}\right)A$$

- **Test d'hypothèse :** $A^* > A^{-1}(1-\alpha)$ Rejeter l'hypothèse H_0 de normalité.

$1-\alpha$	0.75	0.80	0.85	0.90	0.95	0.975	0.99	0.995
$A^{-1}(1-\alpha)$	0.472	0.509	0.561	0.631	0.752	0.873	1.035	1.159

Test de Kolmogorov-Smirnov

Les étapes permettant de réaliser ce test sont :

•Calculer les estimateurs $\bar{x}$ et s à partir des observations $x_1, \ldots, x_n$.

•Ordonner les observations par valeurs croissantes, soit $x_{(1)}, \ldots, x_{(n)}$ l'ensemble ordonné correspondant.

•Calculer les probabilités : $$z_{(k)} = \Phi\left(\frac{x_{(k)} - \bar{x}}{s}\right)$$

Calculer les statistiques : $$D = \max_{k=1\ldots n}\left\{\frac{k}{n} - z_{(k)}, z_{(k)} - \frac{(k-1)}{n}\right\}$$

$$D^* = \left(\sqrt{n} + \frac{0.85}{\sqrt{n}} - 0.01\right)D$$

•**Test d'hypothèse :** $D^* > D^{-1}(1-\alpha)$ Rejeter l'hypothèse H_0 de normalité.

$1-\alpha$	0.85	0.90	0.95	0.975	0.99
$D^{-1}(1-\alpha)$	0.775	0.819	0.895	0.955	1.035

Test Normal Q-plot

Cette méthode graphique s'applique dans le cas où la fonction de répartition F(x) vérifie une équation du type :

$$g\{F(x)\} = a + bh(x)$$

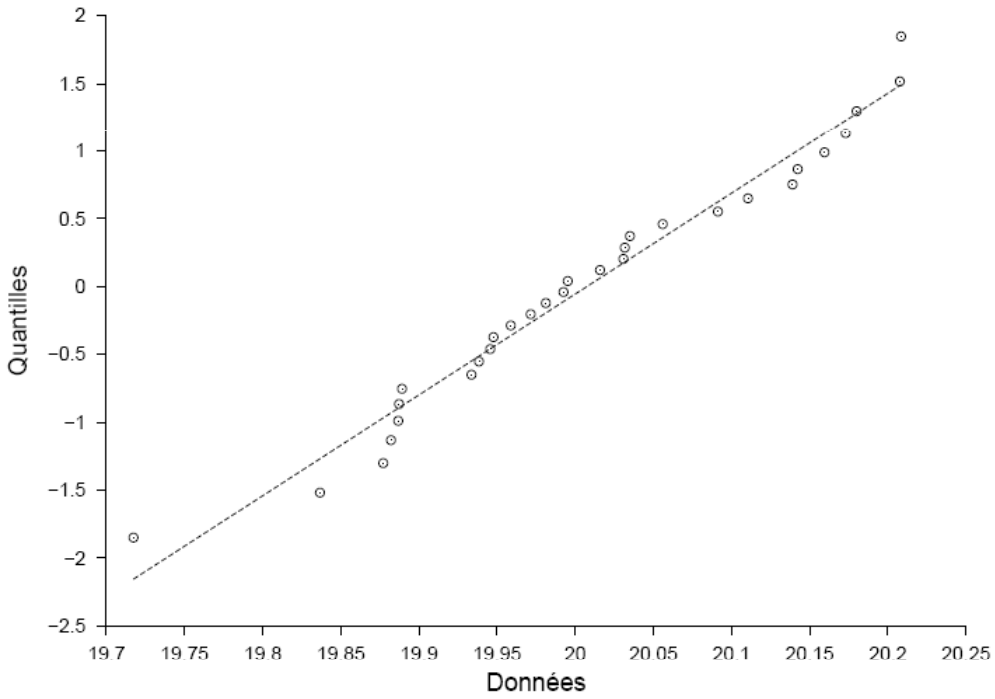
Où *g*() et *h*() sont deux fonctions et où *a* et *b* sont deux paramètres.

Pour tester l'adéquation des observations $x_{(1)}, \ldots, x_{(n)}$ avec la fonction de répartition d'une loi normale, il suffira de tracer les *n* points de coordonnées :

$$\left\{ x_{(k)}, \Phi^{-1}\left(\frac{k}{n+1}\right) \right\}$$

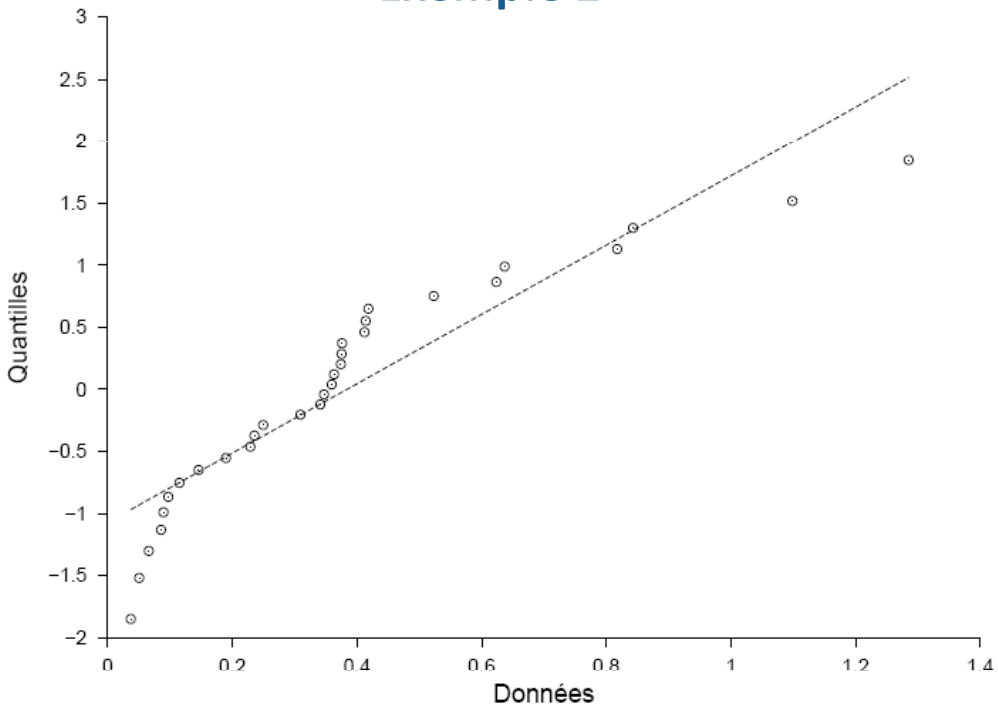
Deux exemples de Q-plot loi normale

Exemple 1



On voit bien que les points sont plutôt alignés, ce qui peut permettre de conclure à la normalité des données

Exemple 2



On voit bien que les points sont moins alignés, donc l'adéquation à une loi normale semble improbable.

Mardia asymétrie et aplatissement

Soit n observations $x_1, \ldots, x_n$ in $\Re^p$:

•Calculer les estimateurs $\bar{x}$ et $\mathbf{S}$ à partir des observations $x_1, \ldots, x_n$.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(x_i - \bar{x}\right)^T$$

•Calculer le coefficient d'asymétrie du test de Mardia :

$$\hat{\delta}_3 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{\left(x_i - \bar{x}\right)^T \mathbf{S}^{-1}\left(x_j - \bar{x}\right)\right\}^3$$

•Calculer le coefficient d'aplatissement du test de Mardia :

$$\hat{\delta}_4 = \frac{1}{n}\sum_{i=1}^{n}\left\{\left(x_i - \bar{x}\right)^T \mathbf{S}^{-1}\left(x_j - \bar{x}\right)\right\}^2$$

Mardia asymétrie et aplatissement

Le test de Hypothèses de ce test de multidimensionnelle de normalité doit être rejeté si :

$$1 - F_{\chi^2}\left\{\frac{n\hat{\delta}_3}{6}, \frac{p(p+1)(p+2)}{6}\right\} \leq \alpha$$

Fonction de répartition do χ^2

or

$$2\Phi\left\{-\frac{\left|\hat{\delta}_4 - p(p+2)\right|}{\sqrt{\frac{8p(p+2)}{n}}}\right\} \leq \alpha$$

Multivarié Normal Q-plot

Soit n observations $x_1, \ldots, x_n$ in $\Re^p$:

Soit x_i un point multidimensionnelle.

$$y_i = (x_i - \bar{x})^T \mathbf{S}^{-1} (x_j - \bar{x})$$

Suit une Fonction de répartition do χ^2 avec p dégrées de liberté

Pour tester l'adéquation des observations x(1), …, x(n) avec la fonction de répartition d'une loi multivarié normale, il suffira de tracer les n points de coordonnées :

$$\left\{ y_i, F_{\chi^2}^{-1}\left(\frac{i}{n+1}, p \right) \right\}$$

Mesures de liaison entre deux variables catégorielles

V de Cramer :

$$\mathbf{V} = \sqrt{\frac{\chi^2}{\chi^2_{\max}}}$$

•mesure directement l'intensité de la liaison de 2 variables catégorielles, sans avoir recours à une table du χ^2

• en intégrant l'effectif et le nombre de degrés de liberté,

par l'intermédiaire de χ^2max

• χ^2max = effectif x [min (nb lignes, nb colonnes) – 1]

• V compris entre 0 (liaison nulle) et 1 (liaison parfaite)

Liaison entre deux variables Continues : Coefficient de corrélation linéaire (Pearson)

Coefficient de Bravais-Pearson: mesure exclusivement le caractère plus ou moins linéaire d'un nuage de points.

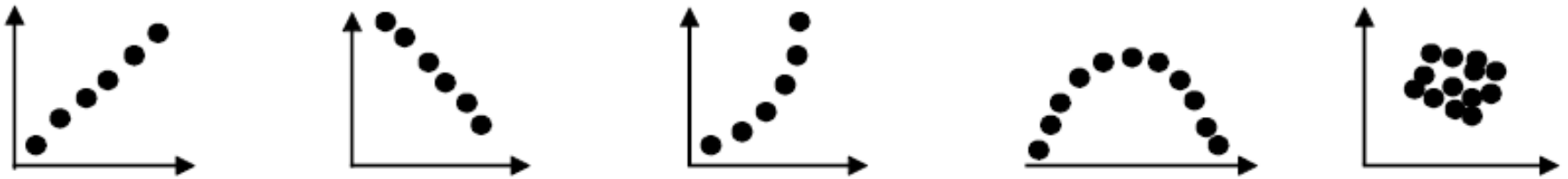
$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \quad \Rightarrow \quad r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- **La liaison est nulle** si le coefficient de corrélation = 0 (nuage de points circulaire ou parallèle à un des 2 axes)
- **La liaison est parfaite** si le coefficient de corrélation = +1 ou -1 (nuage de points rectiligne)
- **La liaison est forte** si le coefficient de corrélation > +0,8 ou < -0,8 (nuage de points elliptique et allongé)

Mais une liaison non linéaire (quadratique) et surtout non monotone n'est pas mesurable par le coefficient de corrélation.

Liaison entre deux variables Continues : Coefficient de corrélation linéaire (Pearson)

Exemples Graphiques:



	A	B	C	D	E
	corrélation positive	corrélation négative	corrélation positive	pas de corrélation, mais dépendance	indépendance
liaison :	monotone linéaire croissante	monotone linéaire décroissante	monotone non linéaire croissante	non monotone	

Liaison entre deux variables Continues : Coefficient de Spearman

Coefficient Rho (ρ) de Spearman plus général car calculé sur les rangs des valeurs et non les valeurs elles-mêmes

- c'est un test non paramétrique (contrairement à Pearson)

- **Préférer le rho de Spearman** si les variables :
  - ne suivent pas une loi normale
  - ont des valeurs extrêmes
  - ne sont pas continues mais ordinales
  - ou pour détecter des liaisons monotones non linéaires

- **Comparer r de Pearson et ρ de Spearman :**
  - r > ρ = présence de valeurs extrêmes
  - ρ > r = liaison non linéaire non détectée par Pearson

Multicolinéarité (corrélation multiple)

- **Concept** : vise les phénomènes d'interdépendance (de corrélation) entre variables explicatives.
- **Multicolinéarité parfaite :**
  - une variable explicative est une combinaison linéaire parfaite des autres variables explicatives.
  - l'estimation des paramètres est impossible, la matrice des données étant singulière.
  - les coefficients du modèle sont indéterminés et leur variance est infinie.
- **Multicolinéarité partielle:**
  - une variable explicative est fortement corrélée à une ou plusieurs variables explicatives (ou à l'une de leur combinaison).
  - les coefficients du modèle de régression peuvent être mais l'écart-type de leur estimation est important.
  - les coefficients ne peuvent donc pas être estimés avec beaucoup de précision.

Multicolinéarité (corrélation multiple)

En théorie, il ne suffit pas de vérifier les variables 2 à 2

- Tolérance d'une variable = proportion de la variance non expliquée par les autres variables - doit être > 0,1
  - VIF (variable inflation factor) = 1 / tolérance
- Indices de conditionnement de la matrice des corrélations
  - multicolinéarité modérée (forte) si des indices ηk > 10 (30)

- **La multicolinéarité est un problème statistique souvent rencontré en régression linéaire.** Elle se manifeste quand certaines colonnes de la matrice **X** sont presque linéairement dépendantes.
- Dans ce cas, la matrice **X'X** est, en termes numériques, tout à fait inversible mais les résultats de la régression sont très instables et en conséquence difficilement interprétables.

Conséquences de la Multicolinéarité

En cas de multicolinéarité parfaite, la sanction est simple :

- l'algorithme des moindres carrés "plante" .

En cas de multicolinéairté classique, on constate en générale :

- que la variance de l'estimation des paramètres tend à être très forte;
- que par conséquent, l'intervalle de confiance autour des paramètres s'élargit considérablement;
- que les tests en t tendent à devenir peu significatifs;
- que malgré cela, le coefficient de détermination peut être très élevé;
- que l'estimation des paramètres est très sensible à la constitution de l'échantillon.

Conséquences de la Multicolinéarité

- **La détection de la présence de multicolinéarité n'est pas toujours évidente.**
  - On peut calculer les coefficients de corrélations entre les variables explicatives prises deux à deux mais cela ne teste que l'existence du phénomène à un niveau d'ordre 1.
  - Le même problème peut se poser au niveau deux (une variable explicative est fortement corrélée avec la combinaison de deux autres variables), au niveau trois, …
- **Une solution alternative** est de réaliser une analyse en composantes principales pour orthogonaliser les variables
  - étudier l'évolution de variance expliquée en fonction du nombre de facteurs
  - mais on perd alors l'interprétabilité directe des coefficients associés aux variables explicatives.

Test ANOVA à 1 facteur

- **Test d'égalité de la moyenne d'une variable continue Y dans k (≥ 2) groupes (définis par les modalités d'une variable nominale)**
  - si plusieurs variables continues dépendantes ⇒ MANOVA
  - si m variables nominales indépendantes ⇒ ANOVA à m facteurs,
- **Généralise le test de Student quand k > 2**
- Ne teste que l'égalité de toutes les moyennes, sans dire le cas échéant lesquelles diffèrent.
- **Exemples :**
  - comparer les productivités de plusieurs usines
  - comparer les rendements de plusieurs champs
  - comparer les effets de plusieurs engrais

Test ANOVA à 1 facteur

•On appelle **analyse de la variance** ce qui est en fait un test d'égalité de la moyenne, en raison de la façon de réaliser ce test, qui consiste à décomposer la variance de la variable continue Y en 2 parties :

- •ce qui peut être attribué aux différences entre groupes (**variance inter-classe**)
- •ce qui peut être attribué aux variations aléatoires (**variance intra-classe, appelée « erreur »**)

Rejeter H_0 :

- •si Carré Moyen inter classe/Carré Moyen intra classe est grand,
- •si les variations aléatoires sont faibles par rapport à l'effet des différences entre classes

Cela se produit quand CMinter/CMintra dépasse la valeur critique de la loi de Fisher au niveau α avec k–1 et n–k degrés de liberté

Modèle Général ANOVA à 1 facteur,

$$\mathbf{Y}_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

- Y_{ij} = valeur de l'obs. j dans le groupe i
- μ = moyenne générale de Y
- α_i = moyenne de Y dans le groupe i – μ
- ε_{ij} = valeur résiduelle

- **distribution normale dans tous les groupes**
  - hypothèse la moins importante pour la qualité du test
- **moyenne = 0**
- **variance égale dans tous les groupes (homoscédasticité)**
- **indépendance ∀ i, j**
  - une observation ne doit pas dépendre des autres du groupe
  - les observations d'un groupe ne doivent pas dépendre de celles des autres groupes
    - cas d'un même individu présent plusieurs fois
    - cas de la comparaison de traitements

Modèle Général ANOVA à 1 facteur,

$$\mathbf{Y}_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

H0 : μ1 = μ2 = … = μk

- les moyennes sont toutes égales
- α1 = α2 = … = αk = 0

H1 : les moyennes ne sont pas toutes égales

- au moins une moyenne est différente
- ne signifie pas : μ1 ≠ μ2 ≠ … ≠ μk
- pour déterminer quelles moyennes diffèrent significativement :
  - test de Bonferroni
  - test de Scheffé (plus puissant)

ANOVA à 1 facteur
Répartition de la somme des carrés

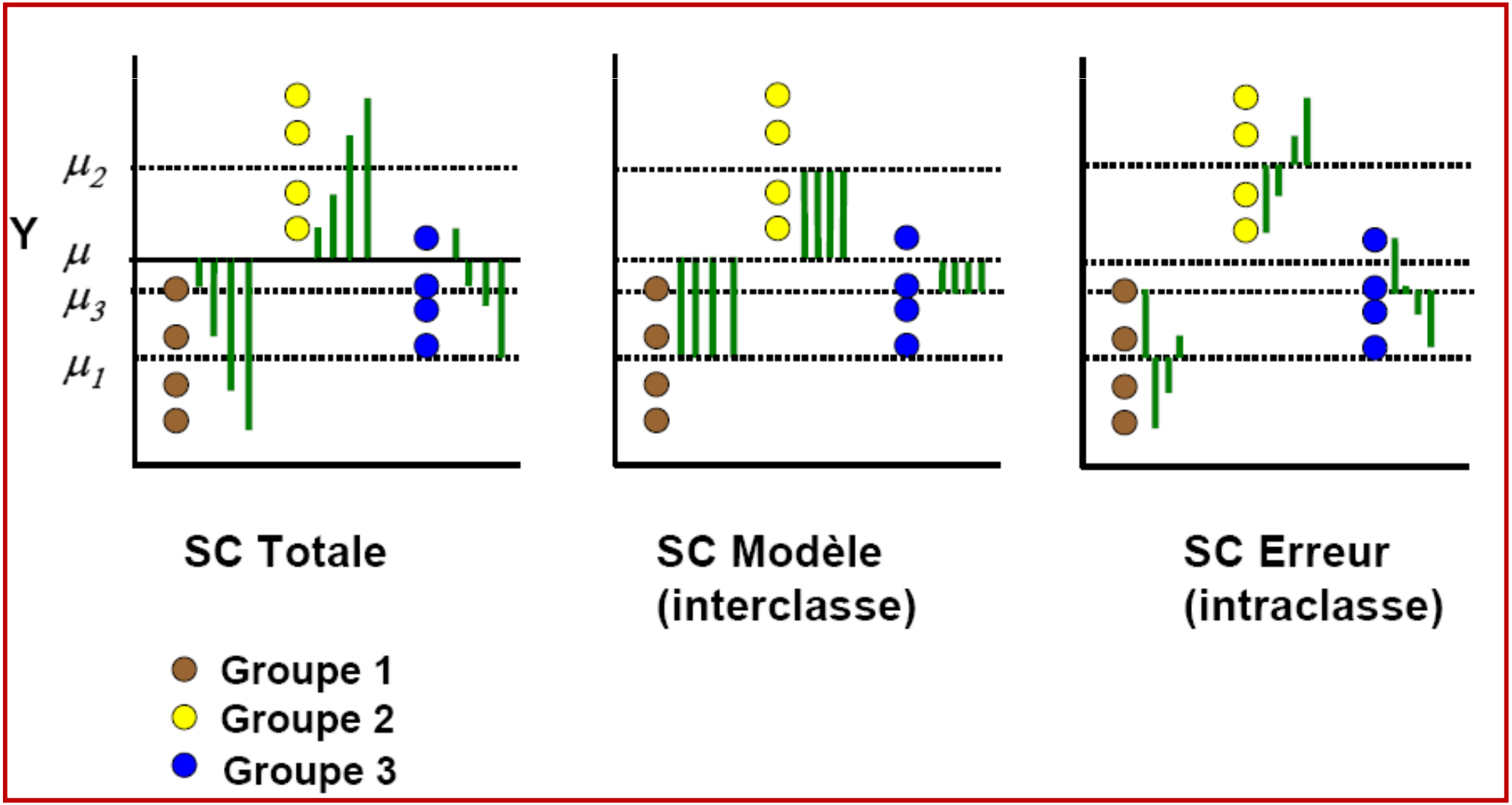


Tableau ANOVA

Source de variation	Somme des carrés (SC)	Degrés de liberté (*dl*)	Carré moyen (CM)	*F*
Totale	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y})^2$	*n - 1*	*SC/dl*	
Inter-classe	$\sum_{i=1}^{k} n_i(\overline{Y}_i-\overline{Y})^2$	*k - 1*	*SC/dl*	$\frac{CM_{interclasse}}{CM_{intraclasse}}$
Intra-classe	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_i)^2$	*n - k*	*SC/dl*	

CM_{inter}/CM_{intra} = F à comparer au F d'une loi de Fisher de ddl (k-1,n-k)

$\eta^2 = SC_{interclasse} / SC_{totale}$ = proportion de la variance expliquée

Estimation de la variance commune σ^2

$$s^2 = \frac{(n_1 - 1)s_1^2 + \ldots + (n_k - 1)s_k^2}{n - k}$$

où $n = n_1 + \ldots + n_k$

s^2 = Within groups Mean-Square

Formules de décomposition

Décomposition de la somme des carrés totale

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2 = \sum_{i=1}^{k} n_i(\bar{y}_i-\bar{y})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ji}-\bar{y}_i)^2$$

Somme des carrés totale (Total)

Somme des carrés inter-classes (Between)

Somme des carrés intra-classes (Within)

Décomposition des degrés de liberté

$$n-1 \quad = \quad (k-1) \quad + \quad (n-k)$$

Mesure de la force de la liaison entre Y et X

Le rapport de corrélation

$$\eta^2 = \frac{\sum_{i=1}^{k} n_i (\bar{y}_i - \bar{y})^2}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ji} - \bar{y})^2} = \frac{\text{Somme des carrés inter-classes}}{\text{Somme des carrés totale}}$$

Test de Comparaison de k moyennes $\mu_1, \mu_2, \ldots, \mu_k$

Test :

$H_0 : \mu_1 = \ldots = \mu_k$

H_1 : Au moins un μ_i différent des autres

Statistique utilisée :

$$F = \frac{\sum_{i=1}^{k} n_i (\bar{y}_i - \bar{y})^2 / (k-1)}{\sum_{i=1}^{k} (n_i - 1) s_i^2 / (n-k)}$$

Règle de décision :

On rejette H_0 au profit de H_1, au risque α de se tromper, si

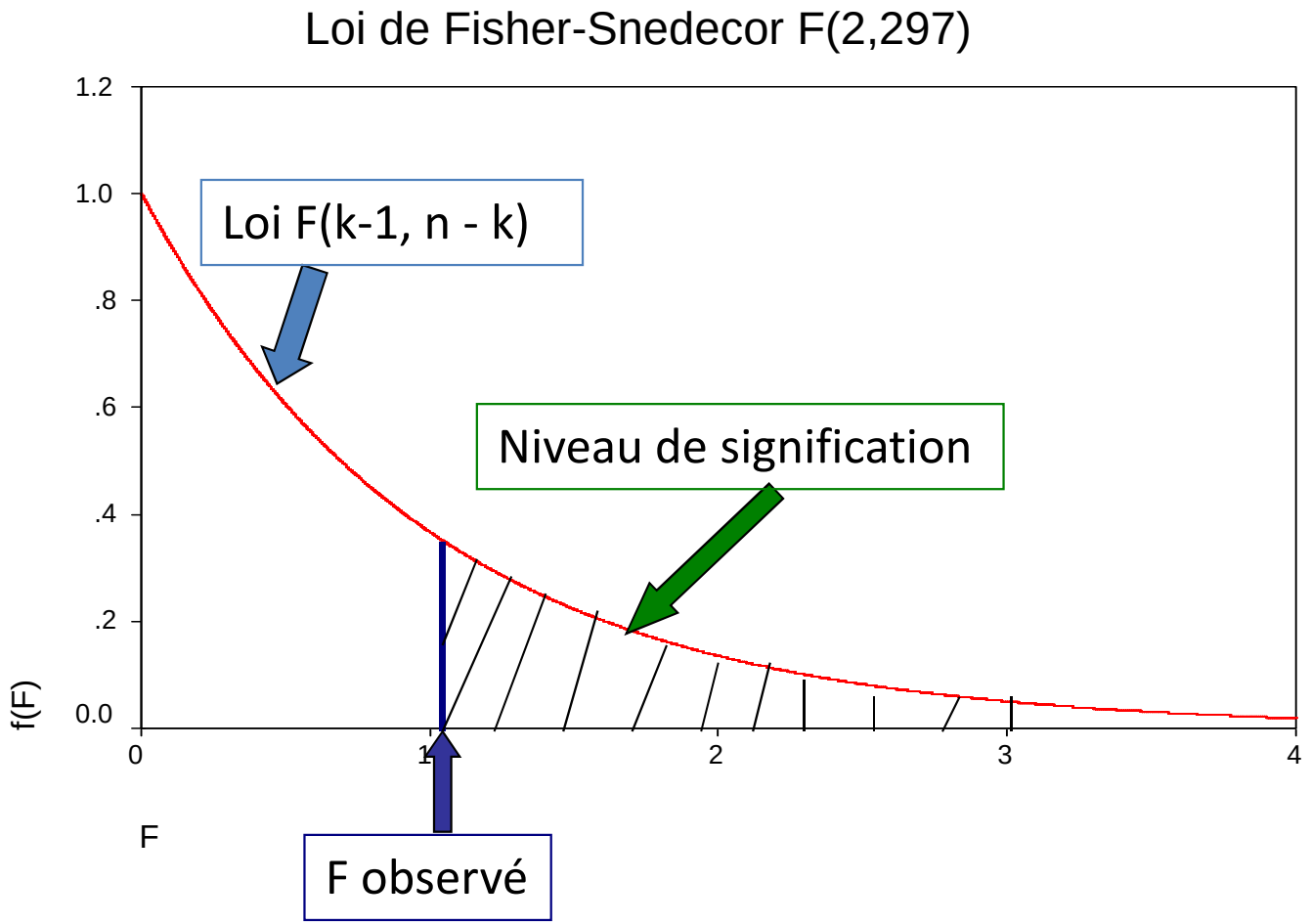
$F \geq F_{1-\alpha}(k-1, n-k)$

Niveau de signification (NS) du F observé :

Plus petite valeur de α conduisant au rejet de H_0 :

$NS = \alpha : F = F_{1-\alpha}(k-1, n-k)$

Niveau de signification du F observé



A ANÁLISE DE VARIÂNCIA (ANOVA)

Formulação matemática do problema:

Modelo Estatístico:

onde:

μ **é a média geral;**

τ_j **é o efeito do grupo j;**

ε_{ij} **é um erro aleatório.**

Hipóteses:

Ho: não há diferenças significativas entre os grupos;

H1: há diferenças significativas entre os grupos.

Exemplo:

Um profissional deseja estudar se a temperatura ambiente influencia na produtividade dos funcionários. Para isso realizou três medidas de produtividade (peças/hora) em três temp. diferentes.

Fator controlável: temperatura

Níveis do fator controlável: 15, 25, 35

Variável de resposta: produtividade

Repetições: 3 valores para cada nível

Temperatura		
15	25	35
12	20	17
13	19	16
11	18	18

Fator controlavel (X)

Niveis do Fator controlavel (xi)

Valores medidos da Variavel resposta Y

Exemplo

	Temperatura			Níveis do controlá
	15°C	25°C	35°C	
	12	20	17	
	13	19	16	
	11	18	18	
$T_{.j} =$	36	57	51	$T_{..} = 144$
$n_j =$	3	3	3	$N = 9$
$\bar{Y}_{.j} =$	12	19	17	$\bar{\bar{Y}}_{..} = 16$

Modelo Estatístico

$$Y_{ij} = \mu + \tau_j + \varepsilon_{ij}$$

20 = 16 + 3 + 1

Decomposição dos resíduos:

$$(Y_{ij} - \overline{\overline{Y}}_{..}) = (\overline{Y}_{.j} - \overline{\overline{Y}}_{..}) + (Y_{ij} - \overline{Y}_{.j})$$

Elevando ao quadrado e somando:

$$\Sigma (Y_{ij} - \overline{\overline{Y}}_{..})^2 = \Sigma n (\overline{Y}_{.j} - \overline{\overline{Y}}_{..})^2 + \Sigma (Y_{ij} - \overline{Y}_{.j})^2$$

$$SQT = SQG + SQR$$

Graus de Liberdade:

$$(N - 1) = (K - 1) + (N - K)$$

Médias quadradas:

$$MQG = SQG / (K - 1)$$
$$MQR = SQR / (N - K)$$

Se não há diferenças significativas entre os grupos

$$E[MQG] = E[MQR]$$

Teste F:

$$F = MQG / MQR$$

Comparar F calculado com F tabelado;

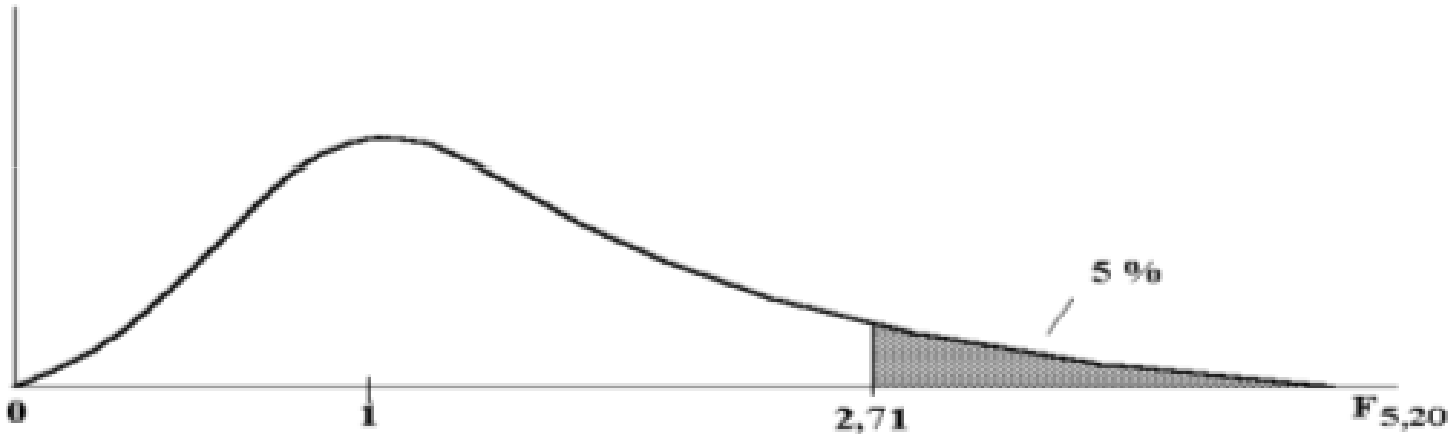
se o valor calculado for maior que o valor tabelado, descarta-se Ho,

ou seja, existe diferenças significativas entre os grupos.

O limite de decisão é estabelecido usando os valores tabelados da distribuição *F*, ou seja:

$$F_{\alpha,k-1,N-k}$$

α : nível de significância
k-1: graus de liberdade do numerador:
N-k : graus de liberdade do denominador:



Distribuição F de Snedecor

Fórmulas para os cálculos:

$$TC = T_{..}^2 / N$$

$$SQT = \Sigma (Y_{ij}^2) - TC$$

$$SQG = \Sigma (T_{.j}^2 / n_j) - TC$$

$$SQR = SQT - SQG$$

Tabela ANOVA:

Fonte	SQ	GDL	MQ	Teste F
Entre Grupos	SQG	K - 1	MQG	F = MQG / MQR
Dentro Grupos	SQR	N - K	MQR	
Total	SQT	N - 1		

Exemplo:

Um pesquisador deseja investigar o efeito da temperatura do forno sobre o número de bactérias contadas após o processo de esterelização. Os dados revelaram o seguinte:

Hipóteses:

Ho: não há diferenças significativas entre os grupos;

H1: há diferenças significativas entre os grupos.

Rappel : Analyse de Variances (ANOVA)

Tempera-tura	70	80	90	100	110	
	15,0	13,1	12,4	10,4	13,1	
	15,9	14,1	11,2	13,4	10,0	
	18,4	18,2	15,9	11,5	13,9	
	17,2	11,1	13,4	14,2	11,1	
	18,6	15,5	9,00	12,7	13,6	
	18,7	12,2	10,3	13,8	12,4	
	16,0	12,3	10,0	12,6	11,2	
	17,1	13,0	13,2	11,4	12,3	
	21,5	15,5	11,0	16,1	13,4	
	14,2	14,3	13,8	13,7	15,9	
	18,4	15,9	12,4	9,20	9,10	
	15,1	15,6	13,4	10,6	10,2	
Totais	206,10	170,80	146,00	149,60	146,20	T..=818,7
No.Obs.	12	12	12	12	12	N = 60
Médias	17,18	14,23	12,17	12,47	12,18	$\bar{\bar{Y}}_{..}$ = 13,65

Cálculos iniciais:

$$TC = T_{..}^2 / N = (818,7)^2 / 60 = 11.171,1$$

$$SQT = \Sigma (Y_{ij}^2) - TC = 11.608,2 - 11.171,1 = 437,1$$

$$SQG = \Sigma (T_{.j}^2 / n_j) - TC$$

$$= [(206,1)^2 / 12] + \ldots + [(146,2)^2 / 12] - 11.171,1 = 222,3$$

$$SQR = SQT - SQG = 437,1 - 222,3 = 214,8$$

Tabela Anova:

Fonte	SQ	GDL	MQ	Teste F
Entre Grupos (Temperatura)	222,3	4	55,6	14,2
Dentro Grupos (Residual)	214,8	55	3,9	
Total	437,1	59		

F calculado	>	F tabelado
14,2	>	2,55

➔ **Há diferenças significativas entre os grupos**

Próximo passo: Comparação múltipla de médias-CMM

1. Calcular o desvio padrão das médias

$$S_{\bar{Y}} = \sqrt{MQR} \; / \sqrt{n_c} = \sqrt{3,9} \; / \sqrt{12} = 1,97 \; / \; 3,46 \; = \; 0,57$$

onde $n_c = (n_1 + n_2 + \ldots + n_k) / k$

2. Calcular o limite de decisão

$$L_d = 3 \times S_{\bar{Y}} = 3 \times 0,57 = 1,71$$

3. Escrever as médias em ordem crescente ou decrescente e compará-las duas a duas. A diferença será significativa se for maior que o L_d

$\overline{Y}_{70} = 17{,}18 \quad \overline{Y}_{80} = 14{,}23 \quad \overline{Y}_{100} = 12{,}47 \quad \overline{Y}_{110} = 12{,}18 \quad \overline{Y}_{90} = 12{,}17$

$\overline{Y}_{70}(17{,}18) - \overline{Y}_{80}(14{,}23) = 2{,}95 > Ld = 1{,}71 \; Dif \; Signif.$

$\overline{Y}_{80}(14{,}23) - \overline{Y}_{100}(12{,}47) = 1{,}76 > Ld = 1{,}71 \; Dif \; Signif.$

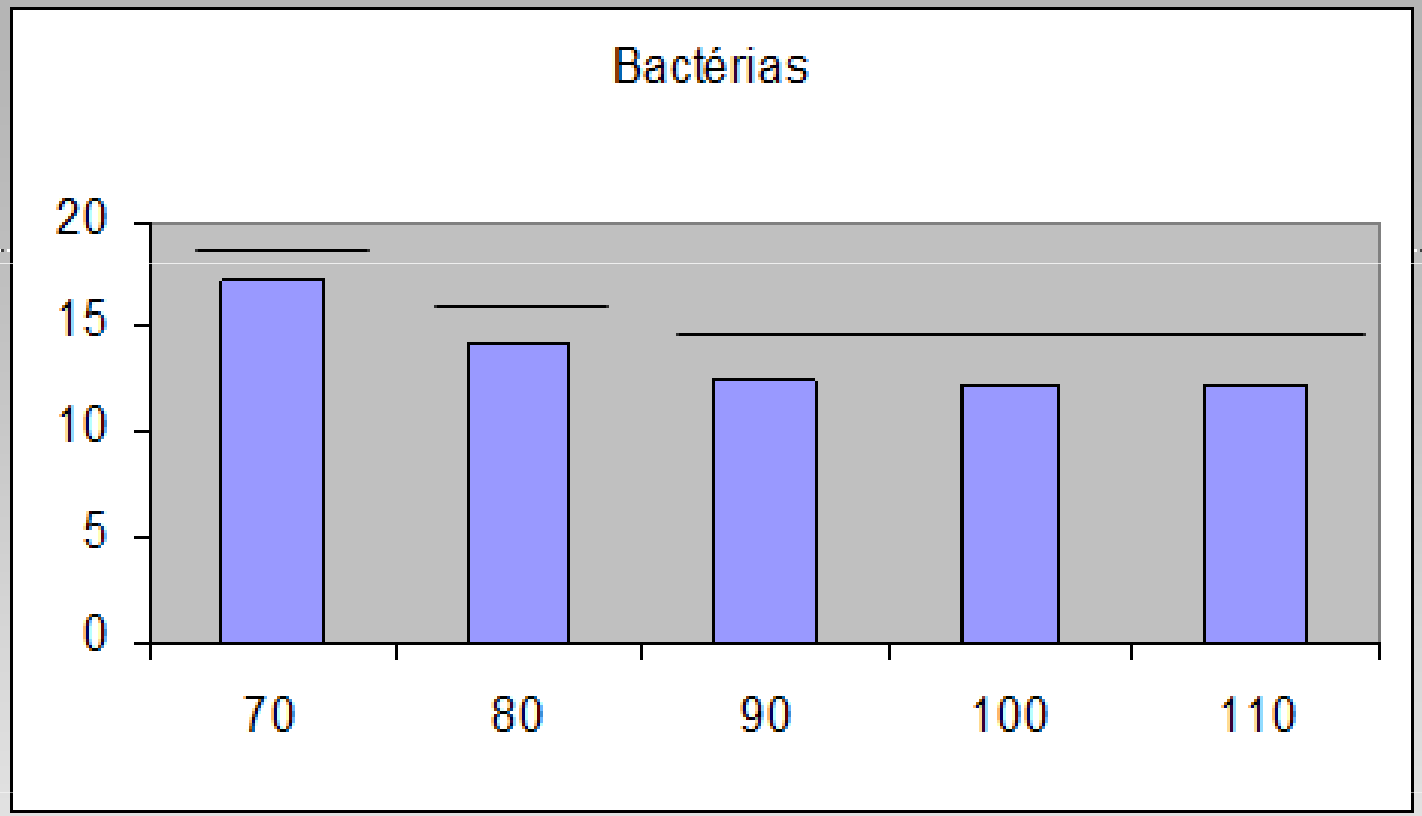
$\overline{Y}_{100}(12{,}47) - \overline{Y}_{110}(12{,}18) = 0{,}29 \; Ld = 1{,}71 \; Dif \; Não \; Signif.$

$\overline{Y}_{110}(12{,}18) - \overline{Y}_{90}(12{,}17) = 0{,}01 \; Ld = 1{,}71 \; Dif \; Não \; Signif.$

4. Usar barras contínuas sobre as médias que não diferem entre si

$\overline{Y}_{70} = 17{,}18 \quad \overline{Y}_{80} = 14{,}23 \quad \overline{Y}_{100} = 12{,}47 \quad \overline{Y}_{110} = 12{,}18 \quad \overline{Y}_{90} = 12{,}17$

Rappel : Analyse de Variances (ANOVA)



- **O ajuste ótimo considerando qualidade (bactérias) é temperatura 90, 100 ou 110**
- **O ajuste ótimo considerando qualidade (bactérias) e custo é temperatura 90 (mais barato)**