

République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITÉ MOHAMED KHIDER, BISKRA
FACULTÉ des SCIENCES EXACTES et des SCIENCES de la NATURE et de la VIE
DÉPARTEMENT INFORMATIQUE



Polycopié du Cours

Analyse de Données

Analyse de Données en Informatique
Première Année Master (M1)

Préparé par :
Dr. AFROUN Fairouz

Université de Biskra, 2023/2024

Table des matières

Table des figures	ii
1 Rappels : Statistiques descriptives & Probabilités	1
Introduction	1
1.1 Rappels sur les statistiques descriptives	1
1.2 Caractérisation d'une variable aléatoire	4
1.2.1 Le concept de variables aléatoires	4
1.2.2 La distribution d'une variable aléatoire	4
1.2.3 La fonction de répartition (distribution cumulative)	4
1.2.4 Espérance (moyenne) d'une variable aléatoire	5
1.2.5 La variance et l'écart-type d'une variable aléatoire	6
1.2.6 Fractals d'une variable aléatoire	6
1.3 Quelques lois de probabilités usuelles	7
1.3.1 La distribution de Gauss (Normale)	7
1.3.2 La distribution χ^2 (<i>Khi – Deux</i>)	8
1.3.3 La distribution Student (<i>t</i>)	9
1.3.4 La distribution Fisher (Fisher-Snedecor) <i>F</i>	9
2 Introduction à la théorie de test d'hypothèses	10
Introduction	10
2.1 Estimateur empirique de la moyenne et de la variance	10
2.1.1 Construction de l'estimateur	10
2.1.2 Distribution de l'estimateur d'une moyenne et d'une variance	11
2.2 Tests de conformité pour une moyenne	12
2.2.1 Cas d'un petit échantillon gaussien ($n \leq 30$ et X de loi normale $N(\mu, \sigma^2)$)	12
2.2.2 Cas d'un grand échantillon : $n > 30$	14
2.3 Tests d'homogénéité	14
2.3.1 Comparaison de deux variances	15
2.3.2 Comparaison de deux moyennes	16
2.4 Analyse de la variance à un facteur (ANOVA 1)	17
2.4.1 Position du problème	18
2.4.2 Analyse de la variance à un seul facteur	18
2.4.3 Les étapes de l'ANOVA 1	19
2.4.4 Exemple d'application	21
Annexe : Tables des lois statistique	22

Table des figures

1.1	Histogramme des effectifs et la courbe des effectifs cumulées	3
1.2	Distribution de Gauss centrée et réduite	7
1.3	Distribution de χ_2 pour différentes degré de liberté.	8
1.4	Distribution de Student.	9

Rappels : Statistiques descriptives & Probabilités

Introduction

L'objectif du présent chapitre est de rappeler quelques notions de base des statistiques descriptives ainsi que de la théorie des probabilités. Plus précisément, dans un premier temps nous allons présenter à travers des exemples numériques les principales caractéristiques descriptives d'une série statistique quantitative (nous se limitons au cas de variables continues) à savoir : les présentations graphiques (Histogramme, polygone des fréquences cumulées,...), les paramètres de position (moyenne, médiane, les quartiles, le mode,...) et les paramètres de dispersions (variance, écart-type,...). Dans un deuxième temps, nous focalisons sur les notions de bases de la théorie des probabilités (variable aléatoire, densité, fonction de répartition, espérance mathématique, etc.). On conclut le chapitre par la présentation de quelques lois usuelles.

1.1 Rappels sur les statistiques descriptives

Supposons qu'on dispose d'une série statistique d'une taille n résumée comme suit :

X	n_i	X_i	$N_i \nearrow$	$F_i \nearrow$	$N_i \searrow$
$[a_0, a_1]$	n_1	$\frac{a_0+a_1}{2}$	$N_1 = 0$	$F_1 = N_1/n$	n
$[a_1, a_2]$	n_2	$\frac{a_1+a_2}{2}$	$N_2 = 0 + n_1$	$F_2 = N_2/n$	$n - n_1$
$[a_2, a_3]$	n_3	$\frac{a_2+a_3}{2}$	$N_3 = 0 + n_1 + n_2$	$F_3 = N_3/n$	$n - n_1 - n_2$
\vdots					
$[a_{i-1}, a_i]$	n_i	$\frac{a_{i-1}+a_i}{2}$	$N_i = 0 + n_1 + \dots + n_i$	$F_i = N_i/n$	$n - n_1 - \dots - n_i$
\vdots					
$[a_{m-1}, a_m]$	n_m	$\frac{a_{m-1}+a_m}{2}$	$N_m = 0 + n_1 + \dots + n_{m-1}$	$F_m = N_m/n$	$n - n_1 - \dots - n_{m-1}$
Σ	n	$-$	n	1	0

alors, la moyenne, la variance et l'écart-type de l'échantillon sont définis respectivement par :

$$\bar{X} = \frac{1}{n} \sum_{i=1}^m n_i X_i; \tag{1.1}$$

$$Var(X) = \frac{1}{n} \sum_{i=1}^m n_i (X_i - \bar{X})^2 = \frac{1}{n} \sum_{i=1}^m n_i X_i^2 - (\bar{X})^2 = \overline{X^2} - (\bar{X})^2. \tag{1.2}$$

$$\acute{E}cart\text{-}type(X) = \sigma(X) = \sqrt{Var(X)}. \tag{1.3}$$

Les différents quartiles de la série en question peuvent être quantifiés à l'aide des formules suivantes :

$$Q_1 = a_i + (a_{i+1} - a_i) \left(\frac{n/4 - N_i}{N_{i+1} - N_i} \right) = a_i + (a_{i+1} - a_i) \left(\frac{1/4 - F_i}{F_{i+1} - F_i} \right); \quad (1.4)$$

tel que a_i et a_{i+1} sont les bornes de la classe qui contient l'élément $X_{n/4}$.

$$Q_2 = Me = a_i + (a_{i+1} - a_i) \left(\frac{n/2 - N_i}{N_{i+1} - N_i} \right) = a_i + (a_{i+1} - a_i) \left(\frac{1/2 - F_i}{F_{i+1} - F_i} \right); \quad (1.5)$$

tel que a_i et a_{i+1} sont les bornes de la classe médiane.

$$Q_3 = a_i + (a_{i+1} - a_i) \left(\frac{n * 3/4 - N_i}{N_{i+1} - N_i} \right) = a_i + (a_{i+1} - a_i) \left(\frac{3/4 - F_i}{F_{i+1} - F_i} \right); \quad (1.6)$$

tel que a_i et a_{i+1} sont les bornes de la classe qui contient $X_{n*3/4}$.

Une notions plus générale est bien que le fractal d'ordre p ($0 \leq p \leq 1$), définis par :

$$Q_p = a_i + (a_{i+1} - a_i) \left(\frac{n * p - N_i}{N_{i+1} - N_i} \right) = a_i + (a_{i+1} - a_i) \left(\frac{p - F_i}{F_{i+1} - F_i} \right); \quad (1.7)$$

tel que a_i et a_{i+1} sont les bornes de la classe qui contient X_{n*p} .

Le mode de la série peut être quantifié en utilisant la formule suivante.

$$M_o = a_i + (a_{i+1} - a_i) \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right); \quad (1.8)$$

tel que : $\Delta_1 = n_i - n_{i-1}$ et $\Delta_2 = n_i - n_{i+1}$ et a_i et a_{i+1} sont les bornes de la classe modale.

Afin de mettre en évidence les différentes notions introduites ci-dessus, nous allons présenter trois exemples numériques.

Exemple 1 Soit le tableau statistique suivant :

X	n_i	X_i	$N_i \nearrow$	$N_i \searrow$	$n_i * X_i$	X_i^2	$n_i * X_i^2$
[144 , 148]	3	146	0	200	438	21316	63948
[148 , 152]	7	150	3	197	1050	22500	157500
[152 , 156]	23	154	10	190	3542	23716	545468
[156 , 160]	32	158	33	167	5056	24964	798848
[160 , 164]	48	162	65	135	7776	26244	1259712
[164 , 168]	41	166	113	87	6806	27556	1129796
[168 , 172]	24	170	154	46	4080	28900	693600
[172 , 176]	15	174	178	22	2610	30276	454140
[176 , 180]	6	178	193	7	1068	31684	190104
[180 , 184]	1	182	199	1	182	33124	33124
Σ	200	-	200	0	32608	-	5326240

Les caractéristiques de cette série sont :

1. La moyenne de la série statistique :

$$\begin{aligned}\bar{X} &= \frac{1}{n} \sum_{i=1}^{10} n_i X_i = \frac{1}{200} (3 * 146 + 7 * 150 + \dots + 1 * 182) \\ &= \frac{1}{200} * 32608 = 163.0400\end{aligned}$$

2. La variance et l'écart-type de la série statistique :

$$\begin{aligned}Var(X) &= \overline{X^2} - (\bar{X})^2 = \frac{1}{n} \sum_{i=1}^{10} n_i X_i^2 - (\bar{X})^2 \\ &= \left(\frac{1}{200} * 5326240 \right) - (163.04)^2 = 49.1584.\end{aligned}$$

et $\sigma(X) = \sqrt{Var(X)} = \sqrt{49.1584} = 7.0113$.

3. Le premier quartile, la médiane et le troisième quartile sont :

On a la classe qui contient $X_{n/4}$ est $[156, 160[$, alors :

$$Q_1 = a_4 + (a_5 - a_4) \left(\frac{200/4 - N_4}{N_5 - N_4} \right) = 156 + (160 - 156) \left(\frac{200/4 - 33}{65 - 33} \right) = 158.1250.$$

On a la classe médiane est $[160, 164[$, alors :

$$Me = a_5 + (a_6 - a_5) \left(\frac{200/2 - N_5}{N_6 - N_5} \right) = 162.9167.$$

On a la classe qui contient $X_{n*3/4}$ est $[164, 168[$, alors :

$$Q_3 = a_6 + (a_7 - a_6) \left(\frac{200*3/4 - N_6}{N_7 - N_6} \right) = 167.6098.$$

4. Le mode : on constate que la classe modale est $[160, 164[$, alors :

$$M_o = a_4 + (a_5 - a_4) \left(\frac{n_5 - n_4}{(n_5 - n_4) + (n_5 - n_6)} \right) = 162.7826.$$

5. L'histogramme et la courbe des effectifs cumulés croissant et décroissant sont présentés dans la figure 1.1.

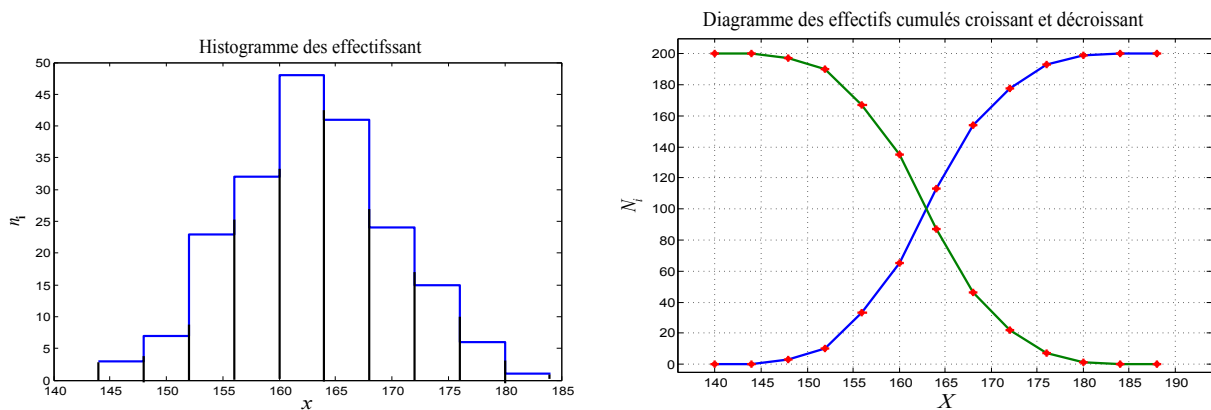


FIGURE 1.1: *Histogramme des effectifs et la courbe des effectifs cumulés*

1.2 Caractérisation d'une variable aléatoire

Le calcul des probabilités s'occupe d'épreuves aléatoires et de phénomènes aléatoires c'est-à-dire d'expériences ou de phénomènes naturels qui, dans des conditions déterminées et stables, ne mènent pas toujours à la même issue. On observe, cependant, une certaine régularité statistique. L'étude de cette régularité fait l'objet d'une théorie mathématique. Dans cette section nous nous limitons à la présentation des outils nécessaires aux applications statistiques qui seront traitées dans les chapitres ultérieurs.

1.2.1 Le concept de variables aléatoires

Une variable aléatoire (*v.a.*) est un nombre réel associé au résultat d'une épreuve, donc un nombre aléatoire. Si l'épreuve est répétée, ce nombre change en général.

Exemple 2 *La taille d'un individu extrait au hasard d'une population ou encore le nombre de "faces" dans une série de 10 jets d'une monnaie.*

1.2.2 La distribution d'une variable aléatoire

1.2.2.1 Cas d'une variable aléatoire discrète

D'une manière générale, si une variable aléatoire discrète X peut prendre les valeurs x_1, x_2, \dots, x_n avec des probabilités respectives p_1, p_2, \dots, p_n , nous dirons que X a pour *distribution de probabilité* l'ensemble des couples :

$$(x_1, p_1), (x_2, p_2), \dots, (x_n, p_n)$$

Notons qu'une distribution de probabilité a les propriétés suivantes :

- $p_i \geq 0$ pour tout i ;
- $\sum_i p_i = 1$.

On peut représenter graphiquement une distribution de probabilité d'une variable aléatoire discrète à l'aide d'un diagramme en bâtons ou en colonnes (comme si c'était une distribution de fréquences dans les statistiques descriptives).

1.2.2.2 Cas d'une variable aléatoire continue

Pour calculer les probabilités afférentes à une variable continue on utilise sa *fonction de densité*, c'est-à-dire une fonction qui permet de calculer la probabilité que X soit dans un intervalle $[a, b]$. Plus précisément, la densité de probabilité (ou densité) de X est une fonction f_X telle que :

1. $f_X(x) \geq 0$ pour tout x ,
2. $\int_{-\infty}^{+\infty} f_X(x) dx = 1$
3. $P(a \leq X \leq b) = \int_a^b f_X(x) dx$.

S'il n'y a pas de possibilité de confusion, on utilisera le symbole f à la place de f_X .

1.2.3 La fonction de répartition (distribution cumulative)

En général, pour exprimer toutes les probabilités associées à une variable aléatoire discrète ou continue il suffit de déterminer les probabilités des intervalles de la forme $I =]-\infty, x]$, où x est un nombre réel. L'outil fondamental qui exprime ces probabilités est la *fonction de répartition*. La *fonction de répartition* d'une variable aléatoire X est la fonction

$$\begin{aligned} F_X(x) &= P(X < x) \\ &= \text{probabilité de l'événement "que } X \text{ soit plus petit à } x\text{".} \end{aligned}$$

Cette fonction est définie pour tout x réel et prend des valeurs entre 0 et 1 ($F_X(x) \in [0, 1]$). S'il n'y a pas de confusion, on utilise le symbole F à la place de F_X .

En général, une fonction de distribution cumulative quelconque a les propriétés suivantes :

- elle est non décroissante,
- elle prend des valeurs entre 0 et 1,
- elle tend vers 0 si x tend vers $-\infty$ et vers 1 si x tend vers $+\infty$.
- La fonction de répartition d'une variable continue est une fonction continue tandis que la fonction de distribution cumulative d'une variable discrète est une fonction discontinue en escalier.
- Entre la distribution de probabilité $(x_1, p_1), (x_2, p_2), \dots$ et la fonction de répartition F d'une variable discrète il y a la relation suivante :

$$F(x) = \sum_{x_i \leq x} p_i.$$

- Entre la fonction de densité f et la fonction de répartition F d'une variable continue, y a les relations suivantes :

$$F(x) = \int_{-\infty}^x f(t) dt; \quad (1.9)$$

$$f(x) = \frac{d}{dx} F(x) \text{ si } F \text{ est dérivable en } x. \quad (1.10)$$

Remarque 1.1 Pour une variable aléatoire continue $P(X \leq x) = P(X < x)$.

1.2.4 Espérance (moyenne) d'une variable aléatoire

Souvent, il suffit d'avoir quelques nombres caractérisant la distribution au lieu de la distribution complète. Les mesures les plus fréquemment utilisées sont la moyenne (ou espérance), la variance, l'écart-type et les fractals.

Soit X une variable aléatoire discrète avec une distribution (x_i, p_i) , $i = 1, 2, \dots, n$. Alors, l'espérance mathématique (ou espérance) de X est définie par :

$$\mu(x) = x_1 p_1 + x_2 p_2 + \dots = \sum_i x_i p_i. \quad (1.11)$$

On utilise aussi le symbole $E(X)$ à la place de $\mu(X)$.

Si X une variable aléatoire continue ayant une densité f alors l'espérance de X est :

$$\mu(x) = \int_{-\infty}^{\infty} x f(x) dx. \quad (1.12)$$

On utilise aussi le symbole $E(X)$ à la place de $\mu(X)$.

Proposition 1.1

1. Une propriété de grande importance est que l'espérance est une application linéaire. Soient X et Y deux variables aléatoires et a et b deux constantes, alors

$$E(aX + bY) = aE(X) + bE(Y); \quad (1.13)$$

2. Une autre propriété très utile, concerne l'espérance d'une transformation d'une variable aléatoire. Soit g une fonction réelle quelconque et $Y = g(X)$ une transformation de X . Alors l'espérance de la variable Y est donnée par :

$$E(Y) = E(g(X)) = \sum_i g(x_i)P(X = x_i), \quad \text{dans le cas discret} \quad (1.14)$$

$$E(Y) = E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)dx, \quad \text{dans le cas continue.} \quad (1.15)$$

1.2.5 La variance et l'écart-type d'une variable aléatoire

L'espérance d'une variable donne une idée de la valeur moyenne de cette variable mais ne prend pas en compte d'autres aspects importants. Par exemple, les variables

- X_1 avec distribution uniforme dans l'intervalle $[-1, 1]$,
- X_2 avec distribution uniforme dans l'intervalle $[-1000, 1000]$,

ont toutes les deux une espérance égale à 0 mais avec une variabilité très différente. Pour mesurer cet aspect on utilise la variance ou l'écart-type.

Soit X une variable aléatoire, la *variance* (de population) de X est définie par :

$$\sigma^2(X) = E([X - E(X)]^2), \quad (1.16)$$

c'est-à-dire,

$$\sigma^2(X) = \sum_i [X_i - E(X)]^2 P(X = x_i), \quad \text{Si } X \text{ discret;} \quad (1.17)$$

$$\sigma^2(X) = \int_{-\infty}^{\infty} [X_i - E(X)]^2 f(x)dx, \quad \text{Si } X \text{ continue.} \quad (1.18)$$

L'écart-type (de population) de X est défini par

$$\sigma(X) = \sqrt{\sigma^2(X)}. \quad (1.19)$$

1.2.6 Fractals d'une variable aléatoire

Le *quantile* (d'ordre) α ($0 < \alpha < 1$) d'une variable aléatoire continue X ayant une fonction de répartition F est le nombre q_α tel que :

$$F(q_\alpha) = \alpha, \quad \text{c'est-à-dire,} \quad q_\alpha = F^{-1}(\alpha). \quad (1.20)$$

Ainsi, on définit des percentiles, des quartiles et des déciles de population. Pour une variable discrète on procède comme dans le cas d'une distribution des fréquences cumulées en statistiques descriptives (voir chapitre 1).

1.3 Quelques lois de probabilités usuelles

Cette partie définit brièvement les modèles de distributions uni-variées les plus fréquemment utilisés comme descriptions approximatives de distributions réelles (en statistique). Comme ces modèles dépendent de paramètres qui doivent être déterminés à l'aide des données que l'on souhaite décrire on les appelle des modèles paramétriques.

1.3.1 La distribution de Gauss (Normale)

On dit que la variable aléatoire X a (ou suit) une distribution normale centrée et réduite ou une distribution de Gauss centrée et réduite qu'on note $X \rightsquigarrow N(0, 1)$ si elle a pour densité

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}. \quad (1.21)$$

Le graphique de f est une courbe "en cloche" (voir 1.2).

Si X a une distribution $N(0, 1)$ on obtient $E(X) = 0$ et $var(X) = 1$.

La fonction de répartition de X est :

$$F(x) = P(X \leq x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt. \quad (1.22)$$

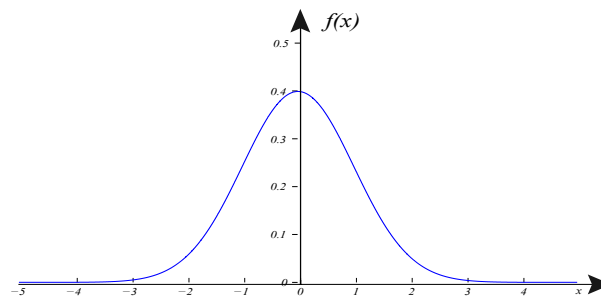


FIGURE 1.2: *Distribution de Gauss centrée et réduite*

Pour déterminer les valeurs de $F(x)$, on se réfère à des tables numériques (voir la Table de la distribution de Gauss) ou on utilise des programmes d'intégration numérique.

Si $X \rightsquigarrow N(0, 1)$ alors, la variable aléatoire $Y = \sigma X + \mu$ a une distribution de Gauss de moyenne μ et de variance σ^2 , notée $N(\mu, \sigma^2)$ et sa densité est donnée par :

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(y-\mu)^2/2\sigma^2}. \quad (1.23)$$

La transformation précédente ($Y = \sigma X + \mu$) effectuée en sens inverse permet de passer d'une variable aléatoire Y de distribution $N(\mu, \sigma^2)$ à la variable centrée et réduite

$$X = \frac{Y - \mu}{\sigma}, \quad (1.24)$$

qui suit une distribution $N(0, 1)$. Cette transformation permet de calculer des probabilités relatives à la variable Y à l'aide de la fonction de répartition et des tables de $N(0, 1)$.

L'un des principaux résultats obtenus sur la distribution Normale est le théorème central limite résumé comme suit :

Théorème 1.1 (Théorème Limite Centrale (TCL))

Supposons que X_1, \dots, X_n soient *i.i.d.* (indépendantes et identiquement distribuées) selon une distribution F_X inconnue, tel que $E(X_i) = \mu$ et $Var(X_i) = \sigma^2$. Alors, si $n \rightarrow \infty$,

$$\frac{\left(\sum_{i=1}^n X_i/n\right) - \mu}{\sigma/\sqrt{n}} \rightsquigarrow N(0, 1). \tag{1.25}$$

Ce théorème peut être interpréter comme suit : La distribution de la moyenne arithmétique centrée et réduite est donc approximativement Gaussienne $N(0, 1)$, indépendamment de la distribution F_X , pourvu que n soit suffisamment élevé. La distribution de la moyenne arithmétique est approximativement normale de moyenne μ et variance σ^2/n et ces paramètres peuvent être estimés. Malheureusement, il n’y a pas en général une règle simple pour déterminer la valeur minimale de n pour que l’approximation soit bonne. Cette valeur dépend de la forme de F_X . Mais généralement, dans la pratique, on se contente de $n \geq 30$.

1.3.2 La distribution χ^2 (Khi – Deux)

Soient X_1, \dots, X_n , n variables aléatoires indépendantes et identiquement distribuées (*i.i.d.*) selon une distribution normale centrée et réduite. On dit que la variable aléatoire

$$Z = X_1^2 + X_2^2 + \dots + X_n^2,$$

à une *distribution χ^2 à n degrés de liberté* notée χ_n^2 . La densité de cette distribution est

$$f(z) = \frac{z^{(n/2)-1}}{2^{n/2}\Gamma(n/2)} e^{-z/2}, \quad z \geq 0, \tag{1.26}$$

avec $\Gamma(\cdot)$ indique la fonction Γ (Gamma), définie par

$$\Gamma(p) = \int_0^\infty x^{p-1} e^{-x} dx, \quad p > 0, \tag{1.27}$$

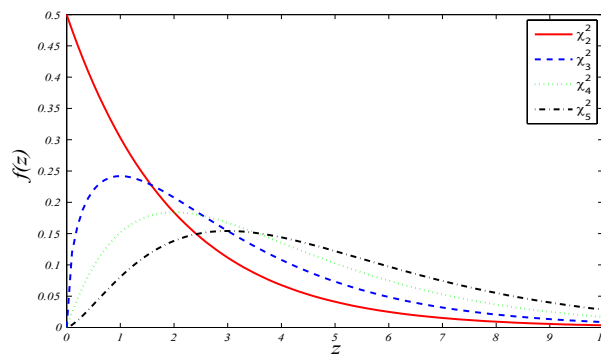


FIGURE 1.3: Distribution de χ^2 pour différentes degré de liberté.

La fonction de répartition est généralement calculée à l’aide d’un programme informatique ou de "tables de la distribution χ^2 " (voir Tables). La moyenne et la variance de la distribution χ^2 sont $E(Z) = n$ et $\sigma^2(Z) = 2n$.

Remarque 1.2 Soit Z_1 et Z_2 deux variables aléatoires de distribution χ^2 de degré liberté n et m respectivement, alors la variable aléatoire $Z = Z_1 + Z_2$ est aussi une variable aléatoire d’une distribution de χ^2 de degré liberté $n + m$ ($Z \rightsquigarrow \chi_{(n+m)}^2$).

1.3.3 La distribution Student (t)

Supposons que X_0, X_1, \dots, X_n , n variables aléatoires indépendantes et identiquement distribuées selon une distribution normale centrée et réduite. On dit que la variable aléatoire

$$T = \frac{X_0}{\sqrt{\frac{1}{n}(X_1^2 + \dots + X_n^2)}} = \frac{X_0}{\sqrt{Z/n}} \quad (\text{avec } Z \rightsquigarrow \chi_n^2) \quad (1.28)$$

à une distribution t (ou distribution de Student) à n degrés de liberté notée t_n . La densité de cette distribution est

$$f(t) = \frac{\Gamma((n+1)/2)}{\Gamma(n/2)\sqrt{n\pi}} (1 + t^2/n)^{-(n+1)/2}. \quad (1.29)$$

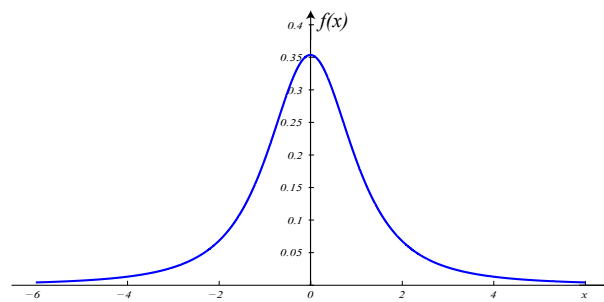


FIGURE 1.4: *Distribution de Student.*

La fonction de distribution cumulative est généralement calculée à l'aide d'un programme informatique ou de "tables de la distribution t " (voir Tables). La moyenne et la variance de la distribution t sont $E(T) = 0$ et $\sigma^2(T) = n/(n-2)$, pour $n > 2$.

1.3.4 La distribution Fisher (Fisher-Snedecor) F

Soit X_1, \dots, X_{n+m} , $n+m$ variables aléatoires indépendantes qui suivent une distribution normale centrée et réduite. On dit que la variable aléatoire

$$Y = \frac{\frac{1}{n}(X_1^2 + \dots + X_n^2)}{\frac{1}{m}(X_{n+1}^2 + \dots + X_{n+m}^2)} = \frac{Z_1/n}{Z_2/m} \quad (Z_1 \rightsquigarrow \chi_n^2 \text{ et } Z_2 \rightsquigarrow \chi_m^2) \quad (1.30)$$

a une distribution F avec n degrés de liberté au numérateur et m degrés de liberté au dénominateur notée $F_{(n,m)}$ ou de degrés de libertés (n, m) . La densité de cette distribution est

$$f(y) = \frac{\Gamma((n+m)/2)}{\Gamma(n/2)\Gamma(m/2)} n^{n/2} m^{m/2} y^{(n/2)-1} (m + ny)^{-(n+m)/2}, \quad y \geq 0. \quad (1.31)$$

Les calculs concernant la distribution F sont généralement effectués à l'aide d'un programme d'ordinateur ou de "tables de la distribution F " (voir Tables). La moyenne de la distribution F est $E(Y) = \frac{m}{m-2}$ pour $m > 2$ et sa variance $\sigma^2(Y) = \frac{2m^2(n+m-2)}{n(m-2)^2(m-4)}$, pour $m > 4$.

Introduction à la théorie de test d'hypothèses

Introduction

Les tests statistiques sont des méthodes de la statistique inférentielle qui, comme l'estimation, permettent d'analyser des données obtenues par tirages au hasard. Ils consistent à généraliser les propriétés constatées sur des observations à la population d'où ces dernières sont extraites, et à répondre à des questions concernant par exemple la nature d'une loi de probabilité, la valeur d'un paramètre ou l'indépendance de deux variables aléatoires.

Il serait important de chercher à présenter en détail l'ensemble des tests statistiques, mais la littérature est très abondante sur le sujet. Pour cela, dans ce chapitre nous allons nous limiter aux tests classiques les plus simples et les plus usuels dans la pratique. En effet, Les tests présentés, sont concernés les tests à un seul échantillon, d'adéquation d'une loi, de comparaison de deux échantillons et enfin l'analyse de la variance à un seul facteur et analyse de la variance à deux facteurs.

2.1 Estimateur empirique de la moyenne et de la variance

2.1.1 Construction de l'estimateur

A partir d'un échantillon (X_1, \dots, X_n) de X , nous définirons la loi de probabilité empirique P_n tel que : $P_n = \frac{1}{n} \sum_{i=1}^{i=n} \delta_{X_i}$ où δ_{X_i} est la masse de Dirac (fonction indicatrice) au point X_i . Cette loi de probabilité admet une fonction de répartition empirique, notée F_n :

$$F_n(x) = \frac{\text{nombre de } x_i \text{ inférieur ou égale à } x}{n} = P_n(]-\infty, x]). \quad (2.1)$$

Définition 2.1 (*moyenne empirique*)

La moyenne empirique d'un n -échantillon (*i.i.d*) est la moyenne de la loi empirique notée :

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{i=n} X_i,$$

cette v.a. admet espérance et variance.

Définition 2.2 (*variance empirique*) La variance empirique de la loi empirique est :

$$\begin{cases} S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2, & \text{si } \mu \text{ est connue;} \\ S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2, & \text{si } \mu \text{ est inconnue.} \end{cases}$$

Remarque 2.1 La valeur moyenne de la variance empirique n'est pas exactement égale à la variance théorique σ^2 (estimateur avec biais), c'est pourquoi on introduit la variance empirique corrigée définie par :

$$S_{n,c}^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

2.1.2 Distribution de l'estimateur d'une moyenne et d'une variance

En général, une simple estimation d'une caractéristique (moyenne, variance, médiane,...) d'un échantillon ne suffit pas : il est nécessaire de connaître son degré d'imprécision. L'outil fondamental pour évaluer un estimateur et le comparer à d'autres, est bien que sa distribution d'échantillonnage. Par exemple, à égalité entre différents aspects, on préférera l'estimateur avec la plus petite variance. Cette section s'occupe de la présentation de la distribution de quelques estimateurs usuels (moyenne, variance).

Considérons un caractère quantitatif représenté par une variable aléatoire X d'espérance mathématique μ , de variance σ^2 , et un échantillon X_1, X_2, \dots, X_n de X de taille n .

1. Pour chaque échantillonnage on peut calculer la moyenne observée du caractère

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i.$$

On démontre que $E(\bar{X}) = \mu$ (un estimateur sans biais de μ) et $Var(\bar{X}) = \frac{\sigma^2}{n}$.

2. Si la moyenne μ est **connue**, alors on considère la variance d'échantillon

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 = \left(\frac{1}{n} \sum_{i=1}^n X_i^2 \right) - \mu^2.$$

3. Si la moyenne μ est **inconnue** alors dans ce cas, l'estimateur sans biais de la variance de l'échantillon est définie comme suit :

$$\hat{\sigma}_c^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{n}{n-1} \hat{S}^2.$$

avec

$$\hat{S}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = \left(\frac{1}{n} \sum_{i=1}^n X_i^2 \right) - \bar{X}^2.$$

Les lois de probabilité de l'estimateur d'une moyenne et d'une variance pour certaine situations peuvent être résumées dans ce qui suit :

1. Cas d'un petit échantillon gaussien $n \leq 30$ et X de loi normale $N(\mu, \sigma^2)$
 - Si σ est connu alors, $U = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ suit la loi normale $N(0, 1)$.
 - Si σ est inconnu, alors, $T = \frac{\bar{X} - \mu}{\frac{\hat{\sigma}_c}{\sqrt{n}}}$ suit la loi de Student à $n - 1$ degrés de liberté.

2. Cas d'un grand échantillon ($n > 30$) et X de loi quelconque :
 - Dans ce cas, $U = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ suit approximativement la loi normale $N(0, 1)$.
3. Cas d'un échantillon gaussien (X de loi normale $N(\mu, \sigma^2)$) :
 - Si μ est connue alors la variable $Y^2 = n \frac{\hat{\sigma}_c^2}{\sigma^2}$ suit la loi de *Khi-Deux* à n degrés de liberté.
 - Si μ est inconnue alors la variable $Y^2 = (n - 1) \frac{\hat{\sigma}_c^2}{\sigma^2}$ suit la loi de *Khi-Deux* à $n - 1$ degrés de liberté.

2.2 Tests de conformité pour une moyenne

Considérons un caractère quantitatif représenté par une variable aléatoire X d'espérance mathématique μ , d'écart-type σ , et un échantillon X_1, X_2, \dots, X_n de taille n de X . La moyenne et la variance corrigée d'échantillon sont données respectivement par :

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \text{ et } \hat{\sigma}_c^2 = \frac{n}{n-1} \hat{\sigma}^2, \text{ avec } \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}^2.$$

2.2.1 Cas d'un petit échantillon gaussien ($n \leq 30$ et X de loi normale $N(\mu, \sigma^2)$)

Dans ce test deux cas sont envisageable. En effet, on peut distinguer le cas où l'écart-type est une quantité bien connue et le cas où l'écart-type n'est connue qu'approximativement à travers son estimateur.

2.2.1.1 Cas σ connu

Il s'agit de faire un choix entre plusieurs hypothèses possibles sur μ sans disposer d'informations suffisantes pour que ce choix soit sûr. On met en avant deux hypothèses privilégiées : l'hypothèse nulle H_0 et l'hypothèse alternative H_1 . Par exemple, on testera

$$H_0 : \text{"}\mu = \mu_0\text{"} \text{ contre } H_1 : \text{"}\mu \neq \mu_0\text{"},$$

avec μ_0 fixé arbitrairement. On veut savoir si l'on doit rejeter H_0 ou pas.

La résolution du présent problème consiste, en résumé, à réaliser les étapes suivantes :

1. Utilise une variable aléatoire dont on connaît la loi de probabilité lorsque H_0 est vraie. Par exemple, on prend $U = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$, en raison que lorsque H_0 est vraie, $U = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$ suit la loi $N(0, 1)$, et cela le fait que l'échantillon est issue d'une variable aléatoire d'une loi normale $X \rightsquigarrow N(\mu, \sigma^2)$.
2. Fixe une valeur $\alpha \in]0, 1[$. En général, on prend α (le risque) petit, le plus souvent

$$\alpha \in \{0.10, 0.05, 0.01, 0.01, 0.001\}.$$

3. Quantifier un réel u_α , tel que $P(-u_\alpha < U < u_\alpha) = 1 - \alpha$. Ce réel u_α peut être extrait de la table de la loi normale centrée et réduite (voir annexe A).
4. Comparer la moyenne empirique \bar{X} de l'échantillon à la moyenne théorique $\mu = \mu_0$, sachant que l'hypothèse H_0 signifiera que les différences observées sont seulement dues aux fluctuations d'échantillonnage (i.e. ne sont pas significatives). En fin, on décide ce qui suit :

- On ne rejettera pas H_0 si les différences observées ne sont pas significatives, c'est-à-dire si U est "petite", ce que l'on peut formuler par $-u_\alpha < U < u_\alpha$, ou encore $|U| < u_\alpha$.
- On rejettera H_0 si les différences observées sont significatives, ce que l'on peut formuler par $U < -u_\alpha$ ou $U > u_\alpha$, c'est-à-dire $|U| > u_\alpha$. Par construction de u_α , on a $P(U > u_\alpha) = P(U < -u_\alpha) = \frac{\alpha}{2}$, soit encore $P(|U| > u_\alpha) = \alpha$ i.e. $P(U \notin] - u_\alpha, u_\alpha[) = \alpha$.

En pratique, on calcule $u = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$ et on décide

- de rejeter H_0 si $u \notin] - \mu_\alpha, \mu_\alpha[$, car si H_0 était vraie, l'événement $U \notin] - \mu_\alpha, \mu_\alpha[$ aurait une probabilité forte de se réaliser ; on pourra dire que la valeur observée \bar{X} n'est pas conforme à la valeur théorique μ_0 mais on ne pourra pas donner de valeurs acceptable de μ ;
- de ne pas rejeter H_0 si $u \in] - \mu_\alpha, \mu_\alpha[$, car si H_0 était vraie, l'événement $U \notin] - \mu_\alpha, \mu_\alpha[$ aurait une probabilité faible de se réaliser ; on pourra dire que la valeur observée \bar{X} est conforme à la valeur théorique μ_0 et que la valeur μ_0 ne peut être rejeter.

Attention : d'autres valeurs μ'_0, μ''_0, \dots peuvent également convenir.

Erreurs de décision Il est à noter que, l'aspect aléatoire de l'échantillon (observations) peut nous faussé la décision finale (rejeter ou non l'hypothèse H_0). En effet, lorsque on rejette H_0 alors que H_0 est vraie, on commet une erreur. On a donc une probabilité α (car lorsque H_0 est vraie, on a $P(U \notin] - \mu_\alpha, \mu_\alpha[) = \alpha$) de se tromper : α est appelée **erreur de première espèce**.

Une autre situation où on peut commettre une erreur de décision est bien que celle lorsque on ne rejette pas H_0 alors que H_0 est fausse. Dans ce cas, on a une probabilité β de se tromper : β est appelée **erreur de deuxième espèce**. Cette probabilité est difficilement calculable car dans la plupart des temps, on ne connaît pas la loi de U lorsque H_0 est fausse. La valeur $1 - \beta$ est appelée la **puissance du test**.

Finalement, ces différentes situations peuvent être résumées par le schéma suivant :

		Réalité	
		H_0	H_1
Décision	H_0	$1 - \alpha$	α
	H_1	β	$1 - \beta$

Les différents tests usuels (formulation et décision) correspondant à la présente situation peuvent être résumés comme suit :

Test (bilatéral) $H_0 : \mu = \mu_0$ contre $H_1 : \mu \neq \mu_0$,

On calcule $u = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$. On détermine u_α , à partir de la table de la loi normale, tel que $P(-u_\alpha < U < u_\alpha) = 1 - \alpha$, et on décide que :

- Si $u \in] - u_\alpha, u_\alpha[$, alors on ne peut rejeter H_0 ;
- Si $u \notin] - u_\alpha, u_\alpha[$, alors on rejette H_0 avec une probabilité α de se tromper.

Test (unilatéral) $H_0 : \mu = \mu_0$ contre $H_1 : \mu > \mu_0$,

On calcule $u = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$. On détermine u_α , à partir de la table de la loi normale, tel que $P(U \geq u_\alpha) = 1 - \alpha$ et on décide que :

- Si $u < u_\alpha$, alors on ne peut rejeter H_0 ;
- Si $u \geq u_\alpha$, alors on rejette H_0 avec une probabilité α de se tromper.

Test (unilatéral) $H_0 : \mu = \mu_0$ contre $H_1 : \mu < \mu_0$,

On calcule $u = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$. On détermine u_α , à partir de la table de la loi normale, tel que $P(U < -u_\alpha) = 1 - \alpha$ et on décide que :

- Si $u > -u_\alpha$, alors on ne peut rejeter H_0 ;
- Si $u \leq -u_\alpha$, alors on rejette H_0 avec une probabilité α de se tromper.

2.2.1.2 Cas σ inconnu

Par définition, on sait que $T = \frac{\bar{X} - \mu_0}{\frac{\hat{\sigma}_c}{\sqrt{n}}}$ suit la loi de Student à $n - 1$ degrés de liberté (voir section 1.3.3). Alors, les différents tests précédents (bilatéral et unilatéral) se font comme suit :

Test (bilatéral) $H_0 : \mu = \mu_0''$ contre $H_1 : \mu \neq \mu_0''$,

On calcule $t = \frac{\bar{X} - \mu_0}{\frac{\hat{\sigma}_c}{\sqrt{n}}}$. On détermine t_α sur la table de Student pour un degré de liberté $n - 1$ tel que $P(-t_\alpha < T < t_\alpha) = 1 - \alpha$ et on décide que :

- Si $t \in] -t_\alpha, t_\alpha[$, alors on ne peut rejeter H_0 ;
- Si $t \notin] -t_\alpha, t_\alpha[$, alors on rejette H_0 avec une probabilité α de se tromper.

Test (unilatéral) $H_0 : \mu = \mu_0''$ contre $H_1 : \mu > \mu_0''$,

On calcule $t = \frac{\bar{X} - \mu_0}{\frac{\hat{\sigma}_c}{\sqrt{n}}}$. On détermine t_α tel que $P(T \geq t_\alpha) = 1 - \alpha$ et on décide que :

- Si $t < t_\alpha$, alors on ne peut rejeter H_0 ;
- Si $t \geq t_\alpha$, alors on rejette H_0 avec une probabilité α de se tromper.

Test (unilatéral) $H_0 : \mu = \mu_0''$ contre $H_1 : \mu < \mu_0''$,

On calcule $t = \frac{\bar{X} - \mu_0}{\frac{\hat{\sigma}_c}{\sqrt{n}}}$. On détermine t_α tel que $P(T < -t_\alpha) = 1 - \alpha$ et on décide que :

- Si $t > -t_\alpha$, alors on ne peut rejeter H_0 ;
- Si $t \leq -t_\alpha$, alors on rejette H_0 avec une probabilité α de se tromper.

2.2.2 Cas d'un grand échantillon : $n > 30$

Dans cette situation ($n > 30$), on se basons sur le TCL, on sait que la variable aléatoire $U = \frac{\bar{X} - \mu_0}{\frac{\hat{\sigma}_c}{\sqrt{n}}}$ suit approximativement une loi normale centrée et réduite ($U \rightsquigarrow N(0, 1)$).

Test (bilatéral) $H_0 : \mu = \mu_0''$ contre $H_1 : \mu \neq \mu_0''$,

On calcule $u = \frac{\bar{X} - \mu_0}{\frac{\hat{\sigma}_c}{\sqrt{n}}}$. On détermine u_α tel que $P(-u_\alpha < U < u_\alpha) = 1 - \alpha$, et on décide que :

- Si $u \in] -u_\alpha, u_\alpha[$, alors on ne peut rejeter H_0 ;
- Si $u \notin] -u_\alpha, u_\alpha[$, alors on rejette H_0 avec une probabilité α de se tromper.

Test (unilatéral) $H_0 : \mu = \mu_0''$ contre $H_1 : \mu > \mu_0''$,

On calcule $u = \frac{\bar{X} - \mu_0}{\frac{\hat{\sigma}_c}{\sqrt{n}}}$. On détermine u_α tel que $P(U \geq u_\alpha) = 1 - \alpha$ et on décide que :

- Si $u < u_\alpha$, alors on ne peut rejeter H_0 ;
- Si $u \geq u_\alpha$, alors on rejette H_0 avec une probabilité α de se tromper.

Test (unilatéral) $H_0 : \mu = \mu_0''$ contre $H_1 : \mu < \mu_0''$,

On calcule $u = \frac{\bar{X} - \mu_0}{\frac{\hat{\sigma}_c}{\sqrt{n}}}$. On détermine u_α tel que $P(U < -u_\alpha) = 1 - \alpha$ et on décide que :

- Si $u > -u_\alpha$, alors on ne peut rejeter H_0 ;
- Si $u \leq -u_\alpha$, alors on rejette H_0 avec une probabilité α de se tromper.

2.3 Tests d'homogénéité

Dans les différents tests présenté dans les sections précédentes on n'a considéré qu'un seule échantillon, pour lequel on s'intéresse si l'un de ses caractères (moyenne, variance, distribution) est conforme à une quantité fixée arbitrairement (cette dernière quantité représente généralement une norme du phénomène étudié). Cependant, dans la pratique, dans certaines situation on dispose de deux populations P_1 et P_2 ou voir même plus de deux populations, dont on étudie un même

caractère et on désire comparer les populations quant à ce caractère, et donc à savoir si elles sont homogènes ou non. Dans cette section, nous se limitons au cas de test d'homogénéité de variance et de moyennes de deux populations indépendantes.

2.3.1 Comparaison de deux variances

Soient X et Y deux variables aléatoires indépendantes représentant le même caractère quantitative dans chacune des populations P_1 et P_2 . On suppose que X et Y suivent des lois normales respectivement, $N(\mu_1; \sigma_1^2)$ et $N(\mu_2; \sigma_2^2)$.

De P_1 , on extrait un échantillon X_1, X_2, \dots, X_{n_1} de taille n_1 de X et de P_2 , on extrait un échantillon Y_1, Y_2, \dots, Y_{n_2} de taille n_2 de Y .

Les moyennes empiriques des deux échantillons sont alors

$$\bar{X} = \frac{1}{n_1} \sum_{i=1}^{n_1} X_i, \quad \bar{Y} = \frac{1}{n_2} \sum_{i=1}^{n_2} Y_i;$$

et leurs variances corrigées sont :

$$\hat{\sigma}_{c,1}^2 = \frac{n_1}{n_1 - 1} \hat{\sigma}_1^2 \text{ avec } \hat{\sigma}_1^2 = \frac{1}{n_1} \sum_{i=1}^{n_1} X_i^2 - \bar{X}^2,$$

$$\hat{\sigma}_{c,2}^2 = \frac{n_2}{n_2 - 1} \hat{\sigma}_2^2 \text{ avec } \hat{\sigma}_2^2 = \frac{1}{n_2} \sum_{i=1}^{n_2} Y_i^2 - \bar{Y}^2.$$

On veut réaliser le test bilatéral suivant :

$$H_0 : \text{''}\sigma_1^2 = \sigma_2^2\text{''} \text{ contre } H_1 : \text{''}\sigma_1^2 \neq \sigma_2^2\text{''}.$$

Les étapes de la réalisation de ce test peuvent être résumées comme suit :

1. On calcule la réalisation $f_c = \frac{\hat{\sigma}_{c,1}^2}{\hat{\sigma}_{c,2}^2}$. Si nécessaire, on permute les échantillons de sorte que $f_c \geq 1$ (c'est-à-dire $f_c = \frac{\max\{\sigma_{c,1}^2, \sigma_{c,2}^2\}}{\min\{\sigma_{c,1}^2, \sigma_{c,2}^2\}}$).
2. Sachant que sous l'hypothèse H_0 , la statistique (variable aléatoire) $F = \frac{\hat{\sigma}_{c,1}^2}{\hat{\sigma}_{c,2}^2}$ suit une loi de Fisher à $(n_1 - 1; n_2 - 1)$ degrés de liberté, alors à partir de la table de Fisher on détermine f_α tel que : $P(F \geq f_\alpha) = \frac{\alpha}{2}$ (ou encore $P(F \leq f_\alpha) = 1 - \frac{\alpha}{2}$).
3. La règle de décision se fait comme suit :
 - si $f_c < f_\alpha$, alors on ne peut rejeter H_0 (H_0 est vraie).
 - si $f_c \geq f_\alpha$, alors on rejette H_0 avec une probabilité α de se tromper.

Avec le même raisonnement on va trouver la zone de non rejet de l'hypothèse nulle dans les tests unilatéral. Les résultats des différents tests sont résumés dans le tableau suivant :

Hypothèse	Zone de non-rejet H_0
$H_0 : \text{''}\sigma_1^2 = \sigma_2^2\text{''} \text{ contre } H_1 : \text{''}\sigma_1^2 \neq \sigma_2^2\text{''}$	$[1; f(n_1 - 1, n_2 - 1, 1 - \frac{\alpha}{2})]$
$H_0 : \text{''}\sigma_1^2 = \sigma_2^2\text{''} \text{ contre } H_1 : \text{''}\sigma_1^2 > \sigma_2^2\text{''}$	$[1; f(n_1 - 1, n_2 - 1, 1 - \alpha)]$
$H_0 : \text{''}\sigma_1^2 = \sigma_2^2\text{''} \text{ contre } H_1 : \text{''}\sigma_1^2 < \sigma_2^2\text{''}$	$[1; f(n_2 - 1, n_1 - 1, 1 - \alpha)]$, avec $f_c = \frac{\hat{\sigma}_{c,2}^2}{\hat{\sigma}_{c,1}^2}$

tel que $f(n, m, 1 - \alpha)$ est lu dans la table de loi Fisher-Snedecor $(1 - \alpha)$ à colonne n , ligne m , de plus on ne rejettera pas H_0 si f_c appartient à la zone de non-rejet de H_0 et on rejettera H_0 sinon.

2.3.2 Comparaison de deux moyennes

Dans cette section, nous allons intéresser à l'homogénéité de deux populations par rapport à la moyenne. Notons que, le test de comparaison de deux moyennes dépend de la distribution des échantillons dont on dispose. Dans le cadre de ce document, nous allons se focalisé sur le cas où les deux échantillons sont de grand taille issues d'une loi quelconque et le cas où les deux échantillons sont gaussien et de petite taille.

2.3.2.1 Cas des grands échantillons

Soient X et Y des variables aléatoires indépendantes représentant le caractère qualitative étudié dans chaque population. On suppose que X et Y suivent une loi quelconque de moyennes respectives μ_1 et μ_2 et d'écart-types respectifs σ_1 et σ_2 . On extrait un échantillon X_1, X_2, \dots, X_{n_1} de taille $n_1 > 30$ de X et un échantillon Y_1, Y_2, \dots, Y_{n_2} de taille $n_2 > 30$ de Y .

Soit la statistique

$$U = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}, \quad (2.2)$$

et u sa réalisation.

Test (bilatéral) $H_0 : \mu_1 = \mu_2$ contre $H_1 : \mu_1 \neq \mu_2$,

Sous l'hypothèse H_0 , la statistique U définie par (2.2) suit approximativement la loi normale centrée réduite $N(0, 1)$.

On calcule $u = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\hat{\sigma}_{c,1}^2}{n_1} + \frac{\hat{\sigma}_{c,2}^2}{n_2}}}$, et on détermine u_α , sur la table de la loi normale, tel que :

$$P(-u_\alpha < U < u_\alpha) = 1 - \alpha,$$

c'est-à-dire

$$P(U < u_\alpha) = 1 - \frac{\alpha}{2},$$

et on décide :

- de ne pas rejeter H_0 si $u \in] -u_\alpha, u_\alpha[$;
- de rejeter H_0 , avec une probabilité α de se tromper, si $u \notin] -u_\alpha, u_\alpha[$.

Test (unilatéral) de $H_0 : \mu_1 = \mu_2$ contre $H_1 : \mu_1 > \mu_2$,

Sous l'hypothèse H_0 , la statistique U suit approximativement la loi normale $N(0, 1)$.

On calcule $u = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\hat{\sigma}_{c,1}^2}{n_1} + \frac{\hat{\sigma}_{c,2}^2}{n_2}}}$, et on détermine u_α , sur la table de la loi normale, tel que :

$P(U \geq u_\alpha) = 1 - \alpha$ et on décide :

- de ne pas rejeter H_0 si $u < u_\alpha$;
- de rejeter H_0 , avec une probabilité α de se tromper, si $u \geq u_\alpha$.

Test (unilatéral) $H_0 : \mu_1 = \mu_2$ contre $H_1 : \mu_1 < \mu_2$,

Sous l'hypothèse H_0 , la statistique U suit approximativement la loi normale $N(0, 1)$.

On calcule $u = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\hat{\sigma}_{c,1}^2}{n_1} + \frac{\hat{\sigma}_{c,2}^2}{n_2}}}$, et on détermine u_α , sur la table de la loi normale, tel que

$P(U < u_\alpha) = 1 - \alpha$ et on décide :

- de ne pas rejeter H_0 si $u > -u_\alpha$;
- de rejeter H_0 , avec une probabilité α de se tromper, si $u \leq -u_\alpha$.

La démarche et les résultats des trois tests ci-dessus restent valable si on remplace σ_1^2 ou σ_2^2 par leurs estimations $\hat{\sigma}_{c,1}^2$, le fait que $U = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}}$ suit aussi une loi normale centrée réduite

(on peut le justifier par le TCL).

2.3.2.2 Cas de petits échantillons ($n \leq 30$)

Soient X et Y des variables aléatoires indépendantes représentant le caractère dans chaque population. On suppose que X et Y suivent une loi normal de moyennes respectives μ_1 et μ_2 , de variance respectives σ_1^2 et σ_2^2 . On extrait un échantillon X_1, X_2, \dots, X_{n_1} de taille $n_1 \leq 30$ de X et un échantillon Y_1, Y_2, \dots, Y_{n_2} de taille $n_2 \leq 30$ de Y .

Test (bilatéral) $H_0 : \mu_1 = \mu_2$ contre $H_1 : \mu_1 \neq \mu_2$,

Afin de réaliser ce test, nous définissons la statistique suivante :

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}}. \tag{2.3}$$

Sous l'hypothèse H_0 et l'hypothèse $\sigma_1 = \sigma_2$ la statistique du test définie dans (2.3) suit approximativement la loi de Student à $n_1 + n_2 - 2$ degrés de liberté.

Cependant, dans la pratique on ne sait pas si $\sigma_1^2 = \sigma_2^2 = \sigma^2$ ou non. A cet effet, on doit d'abord tester l'égalité des deux variances, $\sigma_1^2 = \sigma_2^2$ (Voir section 2.3.1).

Si cette dernière hypothèse est retenue, alors la valeur commune σ^2 peut être estimer par $\hat{\sigma}_c^2 = \frac{(n_1-1)\sigma_{c,1}^2 + (n_2-1)\sigma_{c,2}^2}{n_1+n_2-2}$. Ensuite, on calcule la réalisation de la statistique T , c'est-à-dire $t = \frac{\bar{x} - \bar{y}}{\hat{\sigma}_c \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ et on détermine sur la table de la loi de Student la valeurs critique, t_α , du

test tel que : $P(-t_\alpha < T < t_\alpha) = 1 - \alpha$. Finalement, on décide que :

- On ne peut rejeter H_0 si $t \in] -t_\alpha, t_\alpha[$;
- On rejette H_0 si $t \notin] -t_\alpha, t_\alpha[$, avec une probabilité α de se tromper dans la décision.

Test (unilatéral) $H_0 : \mu_1 = \mu_2$ contre $H_1 : \mu_1 > \mu_2$,

Sous l'hypothèse H_0 , si $\sigma_1 = \sigma_2$ alors la statistique, T , du test définie dans (2.3) suit approximativement la loi de Student à $n_1 + n_2 - 2$ degrés de liberté.

Ainsi, on détermine t_α sur la table de la loi de Student pour un $n = n_1 + n_2 - 2$ et qui vérifié l'égalité $P(T \geq t_\alpha) = 1 - \alpha$ et on décide :

- Si $t < t_\alpha$, alors on ne peut rejeter H_0 ;
- Si $t \geq t_\alpha$, alors on rejette H_0 avec une probabilité α de se tromper dans la décision.

Test (unilatéral) $H_0 : \mu_1 = \mu_2$ contre $H_1 : \mu_1 < \mu_2$,

Sous l'hypothèse H_0 , si $\sigma_1 = \sigma_2$ alors la statistique, T , du test définie dans (2.3) suit encore approximativement la loi de Student à $n_1 + n_2 - 2$ degrés de liberté.

Pour prendre la décision sur le rejet de l'hypothèse H_0 , il suffit de déterminer sur la table de Student pour un ddl $n = n_1 + n_2 - 2$ la valeur critique t_α tel que $P(T < t_\alpha) = 1 - \alpha$ et on décide :

- Si $t > -t_\alpha$, alors on ne peut rejeter H_0 ;
- Si $t \leq -t_\alpha$, alors on rejette H_0 avec une probabilité α de se tromper dans la decision.

2.4 Analyse de la variance à un facteur (ANOVA 1)

Dans cette section, nous allons intéresser à un cas plus générale pour la comparaison de moyennes et cela lorsque le nombre d'échantillon est supérieur strictement à deux. Plus précisément

nous allons intéresser à la technique d'analyse de la variance à un seul facteur qui est la plus adéquate avec la situation.

2.4.1 Position du problème

Supposons que nous ayons 3 bases de données contenant un type d'information bien déterminé où nous désirons savoir si ces bases de données ont une influence sur l'espace mémoire occupé, sur le disque dur, par cette information ou non. À cet effet, nous avons réalisés un recueil de capacité mémoire occupé par six (06) fichiers de chacune des trois bases de données, dont les mesures sont rangées dans le tableau suivant.

N°	Base 1	Base 2	Base 3
1	23.3	18.9	22.5
2	24.4	21.1	22.9
3	24.6	21.1	23.7
4	24.9	22.1	24.0
5	25.0	22.5	24.0
6	26.2	23.5	24.5

TABLE 2.1: *Espace mémoire occupé selon la base de données*

Soit les notions et les notations suivantes :

- La Base de données : Variable qualitative contenant trois modalités, appelée facteur.
- Espace mémoire occupé : Réponse, notée X , et μ_i l'espace mémoire moyen occupé dans la $i^{\text{ème}}$ base de données ($i = \overline{1, 3}$).

Répondre à notre objectif consiste à la réalisation du test suivant :

$$H_0 : "\mu_1 = \mu_2 = \mu_3 = \mu" \text{ contre } H_1 : "\exists i, j \in \{1, 2, 3\} \text{ tel que } \mu_i \neq \mu_j;"$$

Pour réaliser ce test nous pourrions le décomposer en trois sous-tests où nous comparons l'espace mémoire moyen occupé par l'information de l'attribut en question deux à deux selon les bases de données. Mais afin de contourner le problème d'erreur α gonflé, le fait elle ne réalise qu'une seule comparaison à la fois, nous utilisons la technique statistique connue sous le nom d'analyse de variance (en anglais : Analyse Of Variance (ANOVA)) plutôt que des tests de Student t (voir section 2.3.2) multiples. Remarquez que l'ANOVA peut aussi être utilisée quand $p = 2$ puisque, elle retourne la même conclusion qu'un test t .

2.4.2 Analyse de la variance à un seul facteur

L'identification de l'ANOVA 1 au sens littéraire peut être résumée dans la définition suivante :

Définition 2.3 (ANOVA 1)

L'analyse de la variance à un facteur teste l'effet d'un facteur contrôlé A ayant p modalités (groupes) sur les moyennes d'une variable quantitative X .

Les problèmes concernés par la technique ANOVA 1 se présente sous leurs formes générale suivante :

N°	groupe 1	groupe 2		groupe p
1	$X_{1,1}$	$X_{1,2}$	\cdots	$X_{1,p}$
2	$X_{2,1}$	$X_{2,2}$	\cdots	$X_{2,p}$
3	$X_{3,1}$	$X_{3,2}$	\cdots	$X_{3,p}$
4	$X_{4,1}$	$X_{4,2}$	\cdots	$X_{4,p}$
\vdots	\vdots	\vdots	\vdots	\vdots
n_j	$X_{n_1,1}$	$X_{n_2,2}$	\cdots	$X_{n_p,p}$

et le modèle mathématique leurs associés est donné par :

$$X_{ij} = \mu_i + \epsilon_{ij}, \text{ avec } i = \overline{1, n}, j = \overline{1, p} \text{ et } \epsilon_{ij} \rightsquigarrow N(0, \sigma^2), \quad (2.4)$$

où X_{ij} est la $j^{\text{ième}}$ réalisation de la variable quantitative X dans le $i^{\text{ième}}$ échantillon et ϵ_{ij} sont les erreurs de mesure.

Si on retient ce modèle alors le test à réaliser est défini par :

$$H_0 : " \mu_1 = \mu_2 = \dots = \mu_p = \mu " \text{ contre } H_1 : " \exists i, j \in \{1, 2, \dots, p\} \text{ tel que } \mu_i \neq \mu_j ". \quad (2.5)$$

Dans ce qui suit, nous allons énumérer les étapes de la mise en oeuvre de l'ANOVA 1 qui nous permet de réaliser ce test.

2.4.3 Les étapes de l'ANOVA 1

Afin de réaliser le test définie dans (2.5), principalement trois conditions doit être vérifiées préalablement, à savoir :

- Les p échantillons comparés sont indépendants.
- La variable quantitative étudiée suit une loi normale dans les p populations comparées.
- Les p populations comparées ont la même variance : *Homogénéité* des variances ou *homoscédasticité*.

Si ces trois dernières conditions sont vérifiées alors, on peut utiliser la technique ANOVA 1 pour réaliser le test (2.5), et pour ce faire nous avons besoin des quantités (statistiques) suivantes :

- La moyenne de toutes les observations : $\bar{X} = \frac{1}{n} \sum_{j=1}^p \sum_{i=1}^{n_j} X_{ij}$ avec $n = \sum_{j=1}^p n_j$;
- Moyenne de chaque échantillon : $\bar{X}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} X_{ij}$, pour $j = \overline{1, p}$;
- Variance de chaque échantillon : $\hat{\sigma}_i^2 = \frac{1}{n_j} \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)^2$, pour $j = \overline{1, p}$;
- La variance de toutes les observations : $\hat{\sigma}^2 = \frac{1}{n} \sum_{j=1}^p \sum_{i=1}^{n_j} (X_{ij} - \bar{X})^2$ avec $n = \sum_{j=1}^p n_j$.

On peut monter facilement que la variance de toutes les observations est la somme de la variance des moyennes et de la moyenne des variances des p échantillons, c'est-à-dire :

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{j=1}^p \sum_{i=1}^{n_j} (X_{ij} - \bar{X})^2 = \frac{1}{p} \sum_{j=1}^p \sigma_i^2 + \frac{1}{p} \sum_{j=1}^p (\bar{X}_j - \bar{X})^2, \quad (2.6)$$

ou encore :

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{j=1}^p \sum_{i=1}^{n_j} (X_{ij} - \bar{X})^2 = \frac{1}{n} \sum_{j=1}^p \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)^2 + \frac{1}{p} \sum_{j=1}^p (\bar{X}_j - \bar{X})^2. \quad (2.7)$$

On multipliant (2.7), par n on obtient :

$$\underbrace{\sum_{j=1}^p \sum_{i=1}^{n_j} (X_{ij} - \bar{X})^2}_{SC_{Tot}} = \underbrace{\sum_{j=1}^p \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)^2}_{SC_{Res}} + \underbrace{\sum_{j=1}^p \sum_{i=1}^{n_j} (\bar{X}_j - \bar{X})^2}_{SC_{Fac}}, \quad (2.8)$$

où,

SC_{Tot} : est la variation totale qui représente la dispersion des données autour de la moyenne générale.

SC_{Fac} : est la variation due au facteur (variation inter-groupes) qui représente la dispersion des moyennes autour de la moyenne générale.

SC_{Res} : est la variation résiduelle (variation intra-groupes) qui représente la dispersion des données à l'intérieur de chaque échantillon autour de sa moyenne.

L'idée la plus naturelle est de dire que le facteur n'a pas d'impact sur le caractère étudié si la variation totale n'est engendrée que par la variation intra-groupes (résiduelle) associée au caractère, c'est-à-dire,

- Si H_0 est vraie, alors la variation SC_{Fac} due au facteur doit être petite par rapport à la variation résiduelle SC_{Res} .
- Par contre, si H_1 est vraie alors la variation SC_{Fac} due au facteur doit être grande par rapport à la quantité SC_{Res} .

Pour comparer ces quantités, Fisher a considéré le rapport des carrées moyennes associées au facteur CM_{Fac} et les carrées moyennes résiduelles CM_{Res} , où

le carré moyen associé au facteur est : $CM_{Fac} = \frac{SC_{Fac}}{p-1}$.

le carré moyen résiduel est : $CM_{Res} = \frac{SC_{Res}}{n-p}$.

Si les 3 conditions d'application d'ANOVA (Indépendance, Normalité et Homogénéité) sont vérifiées et H_0 est vraie, alors

$$F_{obs} = \frac{CM_{Fac}}{SC_{Res}} \rightsquigarrow f_{(p-1, n-p)}.$$

Décision : Pour un seuil de risque donné α , les tables de Fisher nous fournissent une valeur critique f_α tel que :

$$P\left(\frac{CM_{Fac}}{SC_{Res}} < f_\alpha\right) = 1 - \alpha,$$

- si $f_{obs} < f_\alpha \implies$ on ne peut pas rejeter H_0 (le facteur n'a aucune influence sur le caractère étudié),
- si $f_{obs} \geq f_\alpha \implies$ on rejette H_0 (le facteur influe sur le caractère étudié),
avec f_{obs} est la réalisation de la variable (statistique) F_{obs} .

Les résultats d'une ANOVA 1 sont souvent présentés dans un tableau ayant la forme suivante :

source de variation	Somme des carrés SC	Degrés de libertés ddl	Carré moyen CM	ratio F_{obs}	Ficher c
Inter-groupe (Fac)	SC_{Fac}	$p - 1$	CM_{Fac}	$\frac{CM_{Fac}}{CM_{Res}}$	c
Intra-groupe (Rés)	SC_{Res}	$n - p$	CM_{Res}		
Total	SC_{Tot}	$n - 1$			

2.4.4 Exemple d'application

Reprenant l'exemple présenté dans la section 2.4.1. Les étapes qu'on doit suivre pour réaliser le test

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu \text{ contre } H_1 : \exists i, j \in \{1, 2, 3\} \text{ tel que } \mu_i \neq \mu_j,$$

à l'aide de la technique ANOVA 1, sont les suivantes :

- Calculer les moyennes des différents échantillons : $\bar{X}_1 = 24.73$, $\bar{X}_2 = 21.53$ et $\bar{X}_3 = 23.60$.
- Calculer la moyenne globale de toutes les observations : $\bar{X} = \frac{1}{n}(n_1\bar{X}_1 + n_2\bar{X}_2 + n_3\bar{X}_3) = 23.2889$.
- Compléter le tableau de l'ANOVA à un seul facteur :

source de variation	Somme des carrés <i>SC</i>	Degrés de libertés <i>ddl</i>	Carré moyen <i>CM</i>	ratio F_{obs}	Ficher <i>c</i>
Inter-groupe	31.5911	2	15.7956	12.02	3.6823
Intra-groupe	19.7067	15	1.3138		
Total	51.2978	17			

- Décision : on constate que $f_{obs} = 12.02 > f_{\alpha} = 3.6823$ (pour un risque de $\alpha = 5\%$), donc les espaces moyens occupés par les informations sont significativement différents d'une bases de données à une autre. Cela signifie que le facteur bases de données influe sur l'espace mémoire occupé par les informations stockées.

Conclusion

A partir les différentes notions et différents tests exposés dans ce chapitre on peut conclure que :

Un test d'hypothèse est un procédé d'inférence permettant de contrôler (accepter ou rejeter) à partir de l'étude d'un ou plusieurs échantillons aléatoires, la validité d'hypothèses relatives à une ou plusieurs populations.

Les méthodes de l'inférence statistique nous permettent de déterminer, avec une probabilité donnée, si les différences constatées au niveau des échantillons peuvent être imputables au hasard, ou si elles sont suffisamment importantes pour signifier que les échantillons proviennent de populations vraisemblablement différentes.

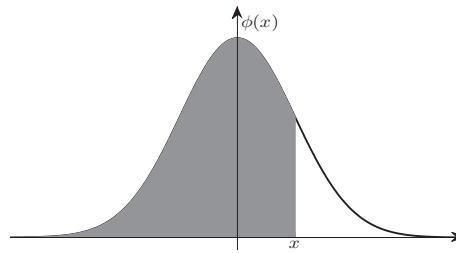
Les tests d'hypothèses font appel à un certain nombre d'hypothèses concernant la nature de la population dont provient l'échantillon étudié (normalité de la variable, égalité des variances, indépendance, etc.) et qui doivent être vérifié préalablement.

Annexe : Tables des lois statistique

Dans cette section on présente les tables les plus représentatives des lois statistiques les plus usuelles. En effet, cette annexe contient les cinq tables suivantes :

1. Table de la loi normale
2. Table de la loi de Student
3. Table de la loi du *Khi – Deux*
4. Table de la loi de Fisher-Snédecor

Fonction de répartition de la loi normale centrée réduite
(probabilité $\phi(x)$ de trouver une valeur inférieure à x)

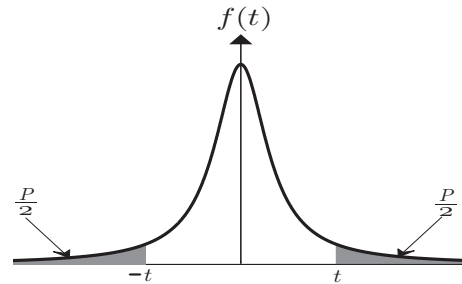


x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.00	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.10	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.20	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.30	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.40	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.50	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.60	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.70	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.80	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.90	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.00	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.10	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.20	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.30	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.40	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.50	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.60	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.70	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.80	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.90	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2.00	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2.10	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.20	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2.30	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.40	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.50	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520
2.60	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.70	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728	0.99736
2.80	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.90	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3.00	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.99900
3.10	0.99903	0.99906	0.99910	0.99913	0.99916	0.99918	0.99921	0.99924	0.99926	0.99929
3.20	0.99931	0.99934	0.99936	0.99938	0.99940	0.99942	0.99944	0.99946	0.99948	0.99950
3.30	0.99952	0.99953	0.99955	0.99957	0.99958	0.99960	0.99961	0.99962	0.99964	0.99965
3.40	0.99966	0.99968	0.99969	0.99970	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976
3.50	0.99977	0.99978	0.99978	0.99979	0.99980	0.99981	0.99981	0.99982	0.99983	0.99983
3.60	0.99984	0.99985	0.99985	0.99986	0.99986	0.99987	0.99987	0.99988	0.99988	0.99989
3.70	0.99989	0.99990	0.99990	0.99990	0.99991	0.99991	0.99992	0.99992	0.99992	0.99992
3.80	0.99993	0.99993	0.99993	0.99994	0.99994	0.99994	0.99994	0.99995	0.99995	0.99995
3.90	0.99995	0.99995	0.99996	0.99996	0.99996	0.99996	0.99996	0.99996	0.99997	0.99997

Table de la loi de Student

Valeurs de T ayant la probabilité P d'être dépassées en valeur absolue

fournit les quantiles t tels que
 $P(|T| \geq t) = p$ pour $t \rightsquigarrow t_n$

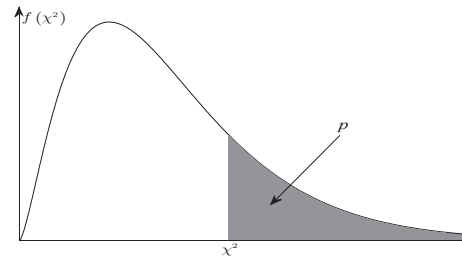


$n \backslash p$	0.90	0.80	0.70	0.60	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01
1	0.158	0.325	0.510	0.727	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66
2	0.142	0.289	0.445	0.617	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925
3	0.137	0.277	0.424	0.584	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841
4	0.134	0.271	0.414	0.569	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604
5	0.132	0.267	0.408	0.559	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032
6	0.131	0.265	0.404	0.553	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707
7	0.130	0.263	0.402	0.549	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499
8	0.130	0.262	0.399	0.546	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355
9	0.129	0.261	0.398	0.543	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250
10	0.129	0.260	0.397	0.542	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169
11	0.129	0.260	0.396	0.540	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106
12	0.128	0.260	0.395	0.539	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055
13	0.128	0.259	0.394	0.538	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012
14	0.128	0.258	0.393	0.537	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977
15	0.128	0.258	0.393	0.536	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947
16	0.128	0.258	0.392	0.535	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921
17	0.128	0.257	0.392	0.534	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898
18	0.127	0.257	0.392	0.534	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878
19	0.127	0.257	0.391	0.533	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861
20	0.127	0.257	0.391	0.533	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845
21	0.127	0.257	0.391	0.532	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831
22	0.127	0.256	0.390	0.532	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819
23	0.127	0.256	0.390	0.532	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807
24	0.127	0.256	0.390	0.531	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797
25	0.127	0.256	0.390	0.531	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787
26	0.127	0.256	0.390	0.531	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779
27	0.127	0.256	0.389	0.531	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771
28	0.127	0.256	0.389	0.530	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763
29	0.127	0.256	0.389	0.530	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756
30	0.127	0.256	0.389	0.530	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750
∞	0.126	0.253	0.385	0.524	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576

- n est le nombre de degrés de liberté.
- Le quantile d'ordre $1 - \frac{\alpha}{2}$ se lit dans la colonne $P = \alpha$.
- Le quantile d'ordre $1 - \alpha$ se lit dans la colonne $P = 2\alpha$.

Table de la loi du *Khi – Deux*
Valeurs de χ^2 ayant la probabilité P d'être dépassées

fournit les quantiles χ^2 tels que
 $P(X \geq \chi^2) = p$ pour $X \rightsquigarrow \chi_n^2$

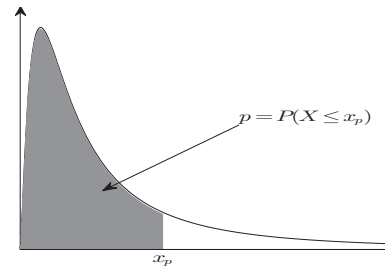


$n \backslash p$	0.995	0.99	0.975	0.95	0.9	0.1	0.05	0.025	0.01	0.005
1	0.00004	0.0002	0.001	0.0039	0.0158	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672

- n est le nombre de degrés de liberté.
- Pour $n > 30$, on peut admettre que la quantité $\sqrt{2\chi^2} - \sqrt{2n-1}$ suit la loi normale centrée réduite.

Tables des quantiles de la *v.a.* de Fisher-Snédecor

fournit les quantiles x_p tels que
 $P(X \leq x_p) = p$ pour $X \rightsquigarrow F_{(n_1; n_2)}$



$$P(X \leq x_p) = 0.95$$

$n_2 \backslash n_1$	1	2	3	4	5	6	7	8	9	10	15	20	30	50	∞
1	161	200	216	225	230	234	237	239	241	242	246	248	250	252	254
2	18.5	19.0	19.2	19.3	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.70	8.66	8.62	8.58	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.86	5.80	5.75	5.70	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.62	4.56	4.50	4.44	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	3.94	3.87	3.81	3.75	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.51	3.44	3.38	3.32	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.22	3.15	3.08	3.02	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.01	2.94	2.86	2.80	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.85	2.77	2.70	2.64	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.72	2.65	2.57	2.51	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.62	2.54	2.47	2.40	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.53	2.46	2.38	2.31	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.46	2.39	2.31	2.24	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.40	2.33	2.25	2.18	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.35	2.28	2.19	2.12	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.31	2.23	2.15	2.08	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.27	2.19	2.11	2.04	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.23	2.16	2.07	2.00	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.20	2.12	2.04	1.97	1.84
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.09	2.01	1.92	1.84	1.71
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.01	1.93	1.84	1.76	1.62
35	4.12	3.27	2.87	2.64	2.49	2.37	2.29	2.22	2.16	2.11	1.96	1.88	1.79	1.70	1.56
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	1.92	1.84	1.74	1.66	1.51
45	4.06	3.20	2.81	2.58	2.42	2.31	2.22	2.15	2.10	2.05	1.89	1.81	1.71	1.63	1.47
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.87	1.78	1.69	1.60	1.44
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.84	1.75	1.65	1.56	1.39
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.02	1.97	1.81	1.72	1.62	1.53	1.35
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00	1.95	1.79	1.70	1.60	1.51	1.32
90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94	1.78	1.69	1.59	1.49	1.30
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.77	1.68	1.57	1.48	1.28
150	3.90	3.06	2.66	2.43	2.27	2.16	2.07	2.00	1.94	1.89	1.73	1.64	1.54	1.44	1.22
200	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93	1.88	1.72	1.62	1.52	1.41	1.19
∞	3.75	2.97	2.59	2.36	2.21	2.09	2.01	1.94	1.88	1.83	1.67	1.57	1.46	1.35	1.00

Tables des quantiles de la *v.a.* de Fisher (suite)

$$P(X \leq x_p) = 0.975$$

$n_2 \backslash n_1$	1	2	3	4	5	6	7	8	9	10	15	20	30	50	∞
1	648	800	864	900	922	937	948	957	963	969	985	993	1001	1008	1018
2	38.5	39.0	39.2	39.2	39.3	39.3	39.4	39.4	39.4	39.4	39.4	39.4	39.5	39.5	39.5
3	17.4	16.0	15.4	15.1	14.9	14.7	14.6	14.5	14.5	14.4	14.3	14.2	14.1	14.0	13.9
4	12.2	10.6	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.66	8.56	8.46	8.38	8.26
5	10.0	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.43	6.33	6.23	6.14	6.02
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.27	5.17	5.07	4.98	4.85
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.57	4.47	4.36	4.28	4.14
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.10	4.00	3.89	3.81	3.67
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.77	3.67	3.56	3.47	3.33
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.52	3.42	3.31	3.22	3.08
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.33	3.23	3.12	3.03	2.88
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.18	3.07	2.96	2.87	2.72
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.05	2.95	2.84	2.74	2.60
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	2.95	2.84	2.73	2.64	2.49
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.86	2.76	2.64	2.55	2.40
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.79	2.68	2.57	2.47	2.32
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.72	2.62	2.50	2.41	2.25
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.67	2.56	2.44	2.35	2.19
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.62	2.51	2.39	2.30	2.13
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.57	2.46	2.35	2.25	2.09
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.41	2.30	2.18	2.08	1.91
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.31	2.20	2.07	1.97	1.79
35	5.48	4.11	3.52	3.18	2.96	2.80	2.68	2.58	2.50	2.44	2.24	2.12	2.00	1.89	1.70
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.18	2.07	1.94	1.83	1.64
45	5.38	4.01	3.42	3.09	2.86	2.70	2.58	2.49	2.41	2.35	2.14	2.03	1.90	1.79	1.59
50	5.34	3.97	3.39	3.05	2.83	2.67	2.55	2.46	2.38	2.32	2.11	1.99	1.87	1.75	1.55
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.06	1.94	1.82	1.70	1.48
70	5.25	3.89	3.31	2.97	2.75	2.59	2.47	2.38	2.30	2.24	2.03	1.91	1.78	1.66	1.44
80	5.22	3.86	3.28	2.95	2.73	2.57	2.45	2.35	2.28	2.21	2.00	1.88	1.75	1.63	1.40
90	5.20	3.84	3.26	2.93	2.71	2.55	2.43	2.34	2.26	2.19	1.98	1.86	1.73	1.61	1.37
100	5.18	3.83	3.25	2.92	2.70	2.54	2.42	2.32	2.24	2.18	1.97	1.85	1.71	1.59	1.35
150	5.13	3.78	3.20	2.87	2.65	2.49	2.37	2.28	2.20	2.13	1.92	1.80	1.67	1.54	1.27
200	5.10	3.76	3.18	2.85	2.63	2.47	2.35	2.26	2.18	2.11	1.90	1.78	1.64	1.51	1.23
∞	4.93	3.67	3.11	2.78	2.56	2.41	2.29	2.19	2.11	2.05	1.83	1.71	1.57	1.43	1.00

$$P(X \leq x_p) = 0.99$$

$n_2 \backslash n_1$	1	2	3	4	5	6	7	8	9	10	15	20	30	50	∞
1	4052	5000	5403	5625	5764	5859	5982	6022	6056	6157	6209	6261	6303	6334	6366
2	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.5	99.5	99.5
3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	26.9	26.7	26.5	26.4	26.1
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.2	14.0	13.8	13.7	13.5
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.72	9.55	9.38	9.24	9.02
6	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.56	7.40	7.23	7.09	6.88
7	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.31	6.16	5.99	5.86	5.65
8	11.3	8.65	7.59	7.01	6.63	6.37	6.19	6.03	5.91	5.81	5.52	5.36	5.20	5.07	4.86
9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	4.96	4.81	4.65	4.52	4.31
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.56	4.41	4.25	4.12	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.25	4.10	3.94	3.81	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.01	3.86	3.70	3.57	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.82	3.66	3.51	3.38	3.17
14	8.86	6.51	5.56	5.04	4.70	4.46	4.28	4.14	4.03	3.94	3.66	3.51	3.35	3.22	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.52	3.37	3.21	3.08	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.41	3.26	3.10	2.97	2.75
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.31	3.16	3.00	2.87	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.23	3.08	2.92	2.78	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.15	3.00	2.84	2.71	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.09	2.94	2.78	2.64	2.42
25	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.22	3.13	2.85	2.70	2.54	2.40	2.17
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.70	2.55	2.39	2.25	2.01
35	7.42	5.27	4.40	3.91	3.59	3.37	3.20	3.07	2.96	2.88	2.60	2.44	2.28	2.14	1.89
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.52	2.37	2.20	2.06	1.80
45	7.23	5.11	4.25	3.77	3.45	3.23	3.07	2.94	2.83	2.74	2.46	2.31	2.14	2.00	1.74
50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.79	2.70	2.42	2.27	2.10	1.95	1.68
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.35	2.20	2.03	1.88	1.60
70	7.01	4.92	4.07	3.60	3.29	3.07	2.91	2.78	2.67	2.59	2.31	2.15	1.98	1.83	1.54
80	6.96	4.88	4.04	3.56	3.26	3.04	2.87	2.74	2.64	2.55	2.27	2.12	1.94	1.79	1.49
90	6.93	4.85	4.01	3.54	3.23	3.01	2.84	2.72	2.61	2.52	2.24	2.09	1.92	1.76	1.46
100	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59	2.50	2.22	2.07	1.89	1.74	1.43
150	6.81	4.75	3.91	3.45	3.14	2.92	2.76	2.63	2.53	2.44	2.16	2.00	1.83	1.66	1.33
200	6.76	4.71	3.88	3.41	3.11	2.89	2.73	2.60	2.50	2.41	2.13	1.97	1.79	1.63	1.28
∞	6.59	4.61	3.79	3.33	3.02	2.81	2.64	2.52	2.41	2.32	2.04	1.88	1.70	1.52	1.00