Cours Statistiques et Analyse de Données

Cours 2 - Partie 2

MASTER INFORMATIQUE PARCOURS SIIA

G. Coppin

2020-2021

Objectifs du cours

Objectifs:

1. Comprendre les mécanismes de l'estimation
2. Savoir estimer une moyenne à partir d'un échantillon
3. Savoir estimer une proportion à partir d'un échantillon

Hypothèse de base

Definition

On supposera que la population étudiée est infinie.

Estimation et statistiques

Le premier travail du statisticien est d'établir un ou plusieurs estimateurs qui décrivent la population à partir de l'échantillon d'observation. Par définition, les estimateurs ne doivent pas dépendre des paramètres réels de la distribution, mais seulement des observables présents dans l'échantillon qui sont supposés être générés à partir de variables aléatoires de même loi P_{θ_0}.

Definition

Un estimateur est toute variable aléatoire construite à partir des observations $X_1, X_2, .., X_n$. En particulier, il ne doit pas dépendre de quantités inconnues, telles que θ_0 ou P_{θ_0}.

Exemple : estimation ponctuelle de proportion

La pratique de l'échantillonnage est incontournable. Si l'échantillon est correctement constitué, la proportion expérimentale devrait être voisine de la proportion théorique. En notant p la proportion réelle (inconnue), n la taille de l'échantillon, et X le nombre d'individus possédant la caractéristique qui nous intéresse (i.e. de **succès**), on peut évaluer:

$$\widehat{p} = \frac{X}{n}$$

$\widehat{p}$ est donc un estimateur de p.

Petit exemple

Deux sondages sont effectués portant sur l'addiction à la Guinness auprès des élèves de Télécom Bretagne. Avec un petit échantillon de 5 individus (rencontrés à 9h du matin le Lundi), on obtient $X = 1$ réponses positives, soit $\widehat{p} = 20\%$. Le même sondage pratiqué auprès de 300 étudiants le Jeudi soir au foyer donne $X = 150$, soit une valeur de $\widehat{p} = 50\%$ mais cette fois ... peut-être plus fiable !

Qualités d'un estimateur

L'estimateur doit être sans biais (ou bien centré) : son espérance est égale à la valeur du paramètre à estimer.

Definition

$$E[\widehat{\theta}] = \theta_0$$

L'estimateur doit être consistant

Definition

$$\theta_n \Rightarrow_{n\to\infty} \theta_0$$

L'estimateur doit avoir une *variance la plus petite possible* pour être aussi *précis* que possible.

Estimation d'une proportion

$\widehat{p}$ est un estimateur sans biais pour la loi binomiale;

$$E(\widehat{p}) = E(\frac{1}{n}X) = \frac{1}{n}E(X) = \frac{1}{n}np = p$$

$$Var(\widehat{p}) = Var(\frac{1}{n}X) = \frac{1}{n^2}Var(X) = \frac{npq}{n^2} = \frac{pq}{n}$$

On retrouve bien l'importance de la taille de l'échantillon pour la qualité (précision) de l'estimation.

Estimation d'une moyenne

De la même façon, on pourra estimer sans biais la moyenne μ d'une loi normale ... par la moyenne des observations dans l'échantillon.

$$E(\overline{X}) = E(\frac{1}{n}\sum_i X_i) = \frac{1}{n}\sum_i E(X_i) = \mu$$

et

$$Var(\overline{X}) = \frac{1}{n^2} Var(\sum_i X_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

Exemple

On suppose que la durée moyenne en minutes de la consommation d'un demi suit une loi normale N, de moyenne inconnue μ et de variance 2. On observe les durées suivantes :

7,3	5,7	6,4	6,7	8,2	6,0	5,8	8,3

La moyenne de ces observations est de $\overline{X} = 6,8$ et la variance de $\overline{X}$ est $\frac{\sigma^2}{n} = \frac{1}{4}$ donc son écart type de $\sigma = 0,5$. Avec 1000 observations, l'écart-type de $\overline{X}$ aurait été de 0,004...

Exemple

On suppose que la durée moyenne en secondes de la consommation d'un demi suit une loi normale N, de moyenne inconnue μ et de variance 2. On observe les durées suivantes :

7,3	5,7	6,4	6,7	8,2	6,0	5,8	8,3

La moyenne de ces observations est de $\overline{X} = 6,8$ et la variance de $\overline{X}$ est $\frac{\sigma^2}{n} = \frac{1}{4}$ donc son écart type de $\sigma = 0,5$. Avec 1000 observations, la variance de $\overline{X}$ aurait été de 0,004...

Intervalle de confiance

- On cherche à placer l'estimation dans un intervalle de confiance qui permette d'apprécier la qualité de l'estimation. Si n est assez grand, l'erreur d'estimation ($\overline{X} - \mu$ ou $\widehat{p} - p$) sera plus petite qu'un écart donné et donc à l'intérieur d'un intervalle.
- En pratique, on définit un risque que l'on accepte de courir, α, qui représente la probabilité que l'intervalle ne contienne pas la véritable valeur du paramètre. $(1 - \alpha)$ est le niveau de confiance de l'intervalle.

$$P(Y_1 < \theta < Y_2) = 1 - \alpha$$

Estimation d'une proportion (I)

$\widehat{p} = \frac{X}{n}$ est un estimateur sans biais de variance $\sigma^2_{\widehat{p}} = \frac{pq}{n}$. Si n grand, grâce au théorême limite central, $\widehat{p}$ est approximativement gaussien de loi $N(E(\widehat{p}) = p, \sigma^2_{\widehat{p}})$, donc $\frac{\widehat{p}-p}{\sigma_{\widehat{p}}}$ est une loi réduite $N(0,1)$. On peut donc trouver un seuil c_α dans la table de loi normale centrée réduite tel que:

$$P(-c_\alpha < \frac{\widehat{p}-p}{\sigma_{\widehat{p}}} < c_\alpha) \simeq 1-\alpha$$

ce qui revient bien à:

$$P(p - c_\alpha \sigma_{\widehat{p}} < \widehat{p} < p + c_\alpha \sigma_{\widehat{p}}) \simeq 1-\alpha$$

ou mieux

$$P(\widehat{p} - c_\alpha \sigma_{\widehat{p}} < p < \widehat{p} + c_\alpha \sigma_{\widehat{p}}) \simeq 1-\alpha$$

Estimation d'une proportion (II)

Il reste à estimer $\sigma_{\widehat{p}} = \sqrt{\frac{pq}{n}}$, ce qui se fait naturellement par $\widehat{\sigma_{\widehat{p}}} = \sqrt{\frac{\widehat{pq}}{n}}$. Ceci permet de déterminer l'intervalle de confiance:

$$(\widehat{p} \pm c_\alpha \widehat{\sigma_{\widehat{p}}}) = (\widehat{p} \pm c_\alpha \sqrt{\frac{\widehat{pq}}{n}})$$

Et là, vous allez me dire ...

Il y en a marre ! Un exemple !

Exemple

- Lors d'un sondage grolandais auprès de 500 personnes, 180 personnes se sont déclarées positives au port de de la casquette à ponpon. Quelle est la proportion théorique p de gens favorables à la casquette à ponpon (avec intervalle de confiance de 90%) ?
- $\widehat{p} = \frac{X}{n} = \frac{180}{500} = 0,360$
- pour avoir $\alpha = 10\%$, on doit prendre $c_\alpha = 1,644854$ (2-tail)
- $(\widehat{p} \pm c_\alpha \sqrt{\frac{\widehat{p}\widehat{q}}{n}}) = (0,360 \pm 1,645 \sqrt{\frac{0,36.0,64}{500}}) = (0,325; 0,395)$

table loi normale

TABLE 3 **Loi Normale $N(0, 1)$: Valeur de $P(N(0,1) > x)$ en fonction de x**

x	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641
0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
1,1	0,1357	0,1335	0,1314	0,1292	0,1971	0,1251	0,1230	0,1210	0,1190	0,1170
1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681
1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
1,8	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233
2,0	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
2,1	0,0179	0,0174	0,0170	0,0166	0,0162	0,0158	0,0154	0,0150	0,0146	0,0143

Remarques méthodologiques et terminologiques

- il faut équilibrer précision et sécurité : plus l'intervalle est petit, plus le risque est grand et réciproquement. Il faut "payer" la précision par un risque d'erreur plus considérable. Pour faire mieux ... il faut augmenter la valeur de n, donc le nombre de sondeurs et le temps de collecte de données.
- on dit souvent que p a 9 chances sur 10 (10% de risque) d'appartenir à l'intervalle de confiance. C'est un abus de langage ... vu que p n'est pas ... (à vous de compléter).

Intervalle vs. risque

Avec $n = 100$ et $\widehat{p} = 0,21$, on peut avoir les intervalles suivants :

α	c_α	Intervalle	longueur
50%	0,674	(0,18 - 0,24)	0,06
10%	1,645	(0,14 - 0,28)	0,14
5%	1,960	(0,13 - 0,29)	0,16
1%	2,576	(0,11 - 0,31)	0,20
0,1%	3,291	(0,08 - 0,34)	0,26

Taille de l'échantillon

Combien doit on prendre d'individus pour que quel que soit p l'intervalle de confiance soit de rayon r_{ref} d'au plus 0,05 (resp. 0,03 / 0,02 / 0,01) ?

- le rayon de l'intervalle de confiance à 95% est égal à $1,960\sqrt{\frac{\widehat{p}_n\widehat{q}_n}{n}}$
- la valeur maximale de $\widehat{p}_n\widehat{q}_n$ est de 0,25 donc $r_{max} = \frac{1,960}{\sqrt{4n}}$
- donc pour avoir $r < r_{max} < r_{ref}$, on doit prendre $n \geq (\frac{0,98}{r_{ref}})^2$
- ce qui donne respectivement $n \geq 385$, $n \geq 1068$, $n \geq 2041$, $n \geq 9604$

Effectivement, la plupart des sondages donnés à 3% sont effectués sur des échantillons de ... plus de 1000 participants.

Application à un sondage d'élection (réel mais ici aux données fictives)

A 20h45, aux élections municipales de 2008, les medias annoncent la victoire de Lyne Cohen-Solal à la mairie du cinquième arrondissement de Paris. A 21h15, ils font marche arrière et annoncent celle de Jean Tiberi. En fait, à 20h45, les statisticiens avaient accès à un échantillon partiel des votes, comme par exemple LCS 473, JT 418 et PM 108. En appliquant l'estimation à chacun des trois candidats, on obtient (à partir du pourcentage empirique mesuré sur les 978 premiers échantillons, avec un risque de 5%) :

- LCS - $[0,473 \pm \frac{1,96\sqrt{0.473*0.527}}{\sqrt{978}}] = [44,2\% - 50,5\%]$
- JT - $[0,418 \pm \frac{1.96\sqrt{0.418*0.582}}{\sqrt{978}}] = [38,7\% - 44,9\%]$
- PM - $[0,108 \pm \frac{1.96\sqrt{0.108*0.892}}{\sqrt{978}}] = [8,8\% - 12,8\%]$

Difficile de s'avancer à 20h45 ...

Estimation d'une moyenne

On a vu que si n grand, la moyenne $\overline{X}$ est approximativement de loi $N(\mu, \sigma^2_{\overline{X}})$. On rappelle que $\sigma^2_{\overline{X}} = \frac{\sigma^2}{n}$. Comme pour les proportions, on va fixer un risque α et poser

$$P(\mu - c_\alpha \sigma_{\overline{X}} < \overline{X} < \mu + c_\alpha \sigma_{\overline{X}}) \simeq 1 - \alpha$$

et de façon identique

$$P(\overline{X} - c_\alpha \sigma_{\overline{X}} < \mu < \overline{X} + c_\alpha \sigma_{\overline{X}}) \simeq 1 - \alpha$$

Estimation d'une moyenne (II)

Comme pour les proportions, il reste à évaluer l'écart type $\sigma_{\overline{X}}$, a priori inconnu. Comme la moyenne est inconnue, on est obligé d'approcher la variance par

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

mais ... cet estimateur est biaisé ! il a une espérance de $\frac{n-1}{n}\sigma^2$. Pour avoir un estimateur sans biais, il faut diviser par $n-1$ et non n soit

$$\widehat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

L'intervalle de confiance de niveau $1-\alpha$ pour μ est alors

$$\overline{X} \pm c_\alpha \frac{\widehat{\sigma}}{\sqrt{n}}$$

Ouf, un autre exemple...

On observe le nombre de plats ingurgités le midi par les utilisateurs d'un restaurant administratif en Bretagne Occidentale, et on obtient les résultats suivants:

Nbre plats	1	2	3	4	5	6	Total
Comptage	230	248	117	76	14	3	688

La suite ...

Le total d'observations est 688, le nombre de plats engloutis est de 1469, ce qui mène à $\overline{X} = 2,135$ et $\widehat{\sigma}^2 = 1,183$ soit $\widehat{\sigma} = 1,088$. Pour avoir un risque de 5%, on doit prendre $c_\alpha = 1,960$ et l'intervalle de confiance est donc :

$$(\overline{X} + c_\alpha \frac{\widehat{\sigma}}{\sqrt{n}}) = (2,135 \pm 1,960.1,088/26,23) = (2,054; 2,216)$$

La suite ...

Le total d'observations est 688, le nombre de plats engloutis est de 1469, ce qui mène à $\overline{X} = 2,135$ et $\widehat{\sigma}^2 = 1,183$ soit $\widehat{\sigma} = 1,088$. Pour avoir un risque de 5%, on doit prendre $c_\alpha = 1,960$ et l'intervalle de confiance est donc :

$$(\overline{X} + c_\alpha \frac{\widehat{\sigma}}{\sqrt{n}}) = (2,135 \pm 1,960.1,088/26,23) = (2,054; 2,216)$$

Observations de loi normale

Si les lois $X_1, X_2, .., X_n$ sont normales, leur moyenne est également normale et $\frac{\overline{X}-\mu}{\sigma_{\overline{X}}}$ est alors exactement de loi $N(0,1)$. Il faut encore traiter le cas $\sigma_{\overline{X}}$ (qui reste inconnu). On le remplace (comme d'habitude) par $\widehat{\sigma}_{\overline{X}}$ et on obtient la variable $\frac{\overline{X}-\mu}{\widehat{\sigma}_{\overline{X}}}$ qui suit ...

Observations de loi normale

Si les lois $X_1, X_2, .., X_n$ sont normales, leur moyenne est également normale et $\frac{\overline{X}-\mu}{\sigma_{\overline{X}}}$ est alors exactement de loi $N(0,1)$. Il faut encore traiter le cas $\sigma_{\overline{X}}$ (qui reste inconnu). On le remplace (comme d'habitude) par $\widehat{\sigma}_{\overline{X}}$ et on obtient la variable $\frac{\overline{X}-\mu}{\widehat{\sigma}_{\overline{X}}}$ qui suit ... une loi de Student !!

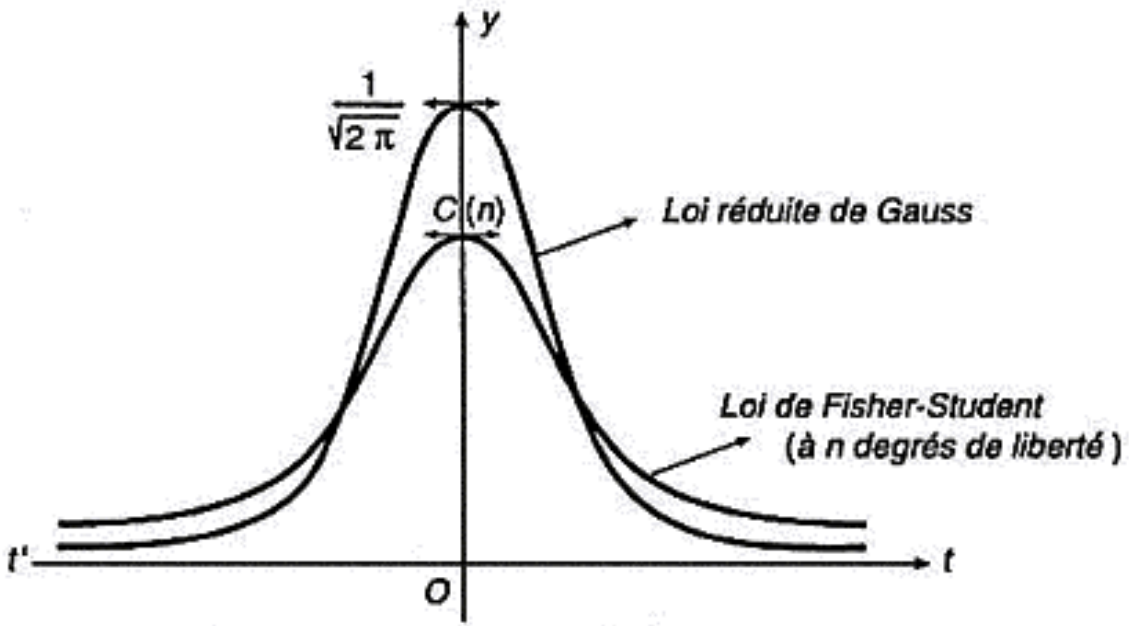
Rappel loi de Student

Soit une variable aléatoire U suivant une loi normale $N(0,1)$ et X indépendante de U suivant une loi χ^2_n. On définit la variable de Student T à n degrés de liberté par:

$$T_n = \frac{U}{\sqrt{\frac{X}{n}}}$$

Dans notre cas, U correspond à la variable normale $\overline{X} - \mu$ et la variable X au dénominateur est bien une variable du χ^2 puisque calculée à partir de la variance (donc somme de gaussiennes). La loi de Student sera paramétrée par un nombre de degrés de liberté égale à $\nu = n - 1$.

Loi normale, loi de Student



Encore et encore un exemple : le problème

On veut estimer la durée moyenne d'une face de disque 33 tours. En mesurant les faces de 5 disques, on obtient le vecteur

$$(17,5-22,4-18,6-24,3-19,5-21,6-15,9-20,4-18,7-20,3)$$

On suppose ces lois normales. Quel est l'intervalle de confiance de niveau 90% pour μ ?

Encore et encore un exemple : la solution

Les observations donnent $\sum X = 199,2$ et $\sum X^2 = 4022,2$. On a donc:

$$\overline{X} = 19,92$$

et

$$\widehat{\sigma}^2 = \frac{\sum X_i^2 - n\overline{X}^2}{n-1} = 5,9951$$

Encore et encore un exemple : la solution

Les observations donnent $\sum X = 199,2$ et $\sum X^2 = 4022,2$. On a donc:

$$\overline{X} = 19,92$$

et

$$\widehat{\sigma}^2 = \frac{\sum X_i^2 - n\overline{X}^2}{n-1} = 5,9951$$

Avec $\alpha = 10\%$ et $\nu = 9$, on obtient $t_\alpha = 1,833$ et l'intervalle de confiance :

$$(\overline{X} \pm t_\alpha \frac{\widehat{\sigma}}{\sqrt{n}}) = (18,50 - 21,34)$$

table loi student

TABLE 4 **Loi de Student t_ν**

Valeur tabulée : argument en fonction de la probabilité et du nombre de degrés de liberté ν.
$P(t_\nu > c) = \alpha$
$\nu = 1(1)30, 40, 60, 120, \infty$

ν \ α	0,25	0,10	0,05	0,025	0,01	0,005	0,0025	0,001	0,0005
1	1,000	3,078	6,314	12,706	31,821	63,657	127,320	318,310	636,620
2	0,816	1,886	2,920	4,303	6,965	9,925	14,089	22,327	31,598
3	0,765	1,638	2,353	3,182	4,451	5,841	7,453	10,214	12,924
4	0,741	1,533	2,132	2,776	3,747	4,604	5,598	7,173	8,610
5	0,727	1,476	2,015	2,571	3,365	4,032	4,773	5,893	6,869
6	0,718	1,440	1,943	2,447	3,143	3,707	4,317	5,208	5,959
7	0,711	1,415	1,895	2,365	2,998	3,499	4,029	4,785	5,408
8	0,706	1,397	1,860	2,306	2,896	3,355	3,833	4,501	5,041
9	0,703	1,383	1,833	2,262	2,821	3,250	3,690	4,297	4,781
10	0,700	1,372	1,812	2,228	2,764	3,169	3,581	4,144	4,587
11	0,697	1,363	1,796	2,201	2,718	3,106	3,497	4,025	4,437
12	0,695	1,356	1,782	2,179	2,681	3,055	3,428	3,930	4,318
13	0,694	1,350	1,771	2,160	2,650	3,012	3,372	3,852	4,221
14	0,692	1,345	1,761	2,145	2,624	2,977	3,326	3,787	4,140
15	0,691	1,341	1,753	2,131	2,602	2,947	3,286	3,733	4,073
16	0,690	1,337	1,746	2,120	2,583	2,921	3,252	3,686	4,015
17	0,689	1,333	1,740	2,110	2,567	2,898	3,222	3,646	3,965
18	0,688	1,330	1,734	2,101	2,552	2,878	3,197	3,610	3,922
19	0,688	1,328	1,729	2,093	2,539	2,861	3,174	3,579	3,883
20	0,687	1,325	1,725	2,086	2,528	2,845	3,153	3,552	3,850
21	0,686	1,323	1,721	2,080	2,518	2,831	3,135	3,527	3,819
22	0,686	1,321	1,717	2,074	2,508	2,819	3,119	3,505	3,792
23	0,685	1,319	1,714	2,069	2,069	2,500	2,807	3,104	3,767
24	0,685	1,318	1,711	2,064	2,492	2,797	3,091	3,467	3,745
25	0,684	1,316	1,708	2,060	2,485	2,787	3,078	3,450	3,725
26	0,684	1,315	1,706	2,056	2,479	2,779	3,067	3,435	3,707
27	0,684	1,314	1,703	2,052	2,473	2,771	3,057	3,421	3,690
28	0,683	1,313	1,701	2,048	2,467	2,763	3,047	3,408	3,674
29	0,683	1,311	1,699	2,045	2,462	2,756	3,038	3,396	3,659
30	0,683	1,310	1,697	2,042	2,457	2,750	3,030	3,385	3,646
40	0,681	1,303	1,684	2,021	2,423	2,704	2,971	3,307	3,551
60	0,679	1,296	1,671	2,000	2,390	2,660	2,915	3,232	3,460
120	0,677	1,289	1,658	1,980	2,358	2,617	2,860	3,160	3,373

Estimation d'un paramètre quelconque

Pour un paramètre quelconque θ pour une loi quelconque, on procède de manière analogue :

- on cherche un estimateur convenable $\widehat{\theta}$ dont la variance peut être estimée $\sigma^2_{\widehat{\theta}}$ - cette estimation se fait souvent en remplaçant dans la formule de la variance θ par $\widehat{\theta}$.
- pour n grand, $\widehat{\theta}$ se comportera comme une loi normale et la formule générale $(\widehat{\theta} \pm c_\alpha \sigma_{\widehat{\theta}})$ donnera l'intervalle de confiance

Petit retour en arrière

- On applique les estimateurs précédents sans hésiter ... lorsque la population est infinie (vraiment infinie, avec remise, grande devant la taille de l'échantillon)
- Lorsque la population est finie, les estimateurs de la variance changent !

L'estimateur de la variance est, pour une taille de population N et une taille d'échantillon n

$$\widehat{\sigma}_{\overline{X}} = \frac{\widehat{\sigma}}{\sqrt{n}}\sqrt{1-\frac{n}{N}}$$

Intervalle de confiance avec population finie

L'intervalle de confiance de niveau $1-\alpha$ pour μ est alors

$$\overline{X} \pm c_\alpha \frac{\widehat{\sigma}}{\sqrt{n}} \sqrt{1-\frac{n}{N}}$$