Echantillonnage et inférence

Introduction

On s'intéresse désormais à l'inférence.

On note par

$$\{X_i\}_{i=1}^{n} = \{X_1, X_2, \ldots, X_n\},$$

un échantillon de n observation de la variable X, tirée de $f_X(\cdot)$ (la population).

L'échantillon nous permet d'en apprendre plus sur la population.

L'échantillon est une random variable (!) (potentiellement multivariée)

Ici , nous illustrerons les concepts avec des ensembles finis mais il faut également se pencher sur la théorie asymptotique.

Les estimateurs et les tests seront abordés ensuite.

L'échantillonnage aléatoire

Définition

$\{X_i\}_{i=1}^{n}$ est un échantillon de taille n de f_X i :

(i) les X_i sont mutuellement indépendants;

(ii) $f_{X_i} = f_X$ pour tout i.

Souvent on dira que $X_1, \cdots, X_n$ sont "independent et identically distributed"

Avec pour densité $f_X(x)$. [iid]

De pl

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i) = \prod_{i=1}^{n} f_X(x_i)$$

Exemples

Si $X \sim N(\mu, \sigma^2)$, avec $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ Alors :

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2} = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \prod_{i=1}^{n} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}$$
$$= (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2}.$$

Si $X \sim$ Bernoulli(π), avec $f(x) = \pi^x (1-\pi)^{1-x}$ alors

Avec: $\bar{x} = n^{-1} \sum_{i=1}^{n} x_i$ et $\sum_{i=1}^{n}(1-x_i) = n - \sum_{i=1}^{n} x_i = n - n\bar{x}.$

Défnition

Une statistique d'échantillon est une fonction de X telle que, $\theta_n = \theta(X_1, \ldots, X_n)$.

Définition très larges avec pour seule restriction que θ ne peut pas être une fonction d'un paramètre .

Puisque l'échantillon est aléatoire, la statistique aussi. En conséquence, θ_n a une distribution.

Exemples

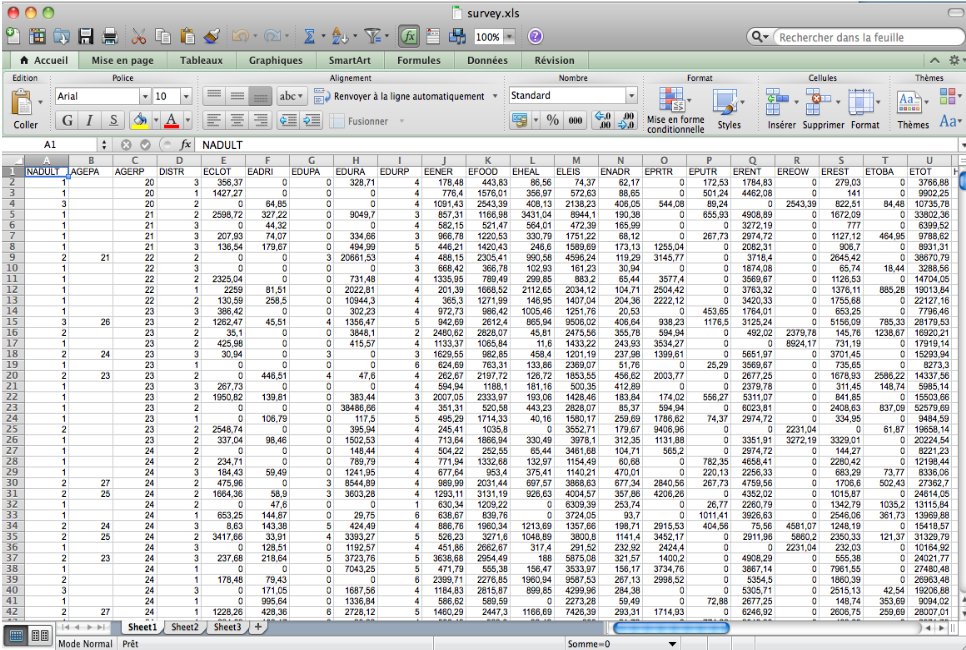
Moyenne d'échantillon : $\overline{X} = n^{-1} \sum_{i=1}^{n} X_i$;

Variance d'échantillon : $S_X^2 = (n-1)^{-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$;

Écart type de l'échantillon : $\sqrt{S_X^2} = S_X$.

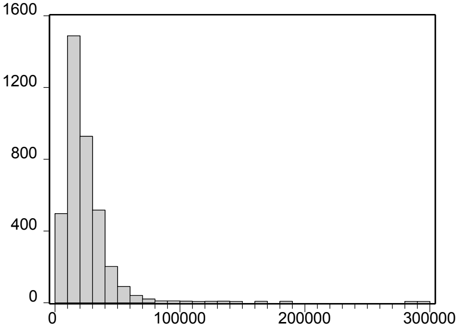
On Utilise n-1 pour la variance comme correction de degrés de liberté

Matrice de donnée d'un sondage de budget



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	NADULT	AGEPA	AGERP	DISTR	ECLOT	EADRI	EDUPA	EDURA	EDURP	EENER	EFOOD	EHEAL	ELEIS	ENADR	EPRTR	EPUTR	ERENT	EREOW	EREST	ETOBA	ETOT
2	1		20	3	356,37	0	0	328,71	4	178,48	443,83	86,56	74,37	62,17	0	172,53	1784,83	0	279,03	0	3766,88
3	1		20	1	1427,27	0	0	0	4	776,4	1576,01	356,97	572,63	88,65	0	501,24	4462,08	0	141	0	9902,25
4	3		20	2	0	64,85	0	0	4	1091,43	2543,39	408,13	2138,23	406,05	544,08	89,24	0	2543,39	822,51	84,48	10735,78
5	1		21	2	2598,72	327,22	0	9049,7	3	857,31	1166,98	3431,04	8944,1	190,38	0	655,93	4908,89	0	1672,09	0	33802,36
6	1		21	3	0	44,32	0	0	4	582,15	521,47	564,01	472,39	165,99	0	0	3272,19	0	777	0	6399,52
7	1		21	3	207,93	74,07	0	334,66	3	966,78	1220,53	330,79	1751,22	68,12	0	267,73	2974,72	0	1127,12	464,95	9788,62
8	1		21	3	136,54	179,67	0	494,99	5	446,21	1420,43	246,6	1589,69	173,13	1255,04	0	2082,31	0	906,7	0	8931,31
9	2	21	22	2	0	0	3	20661,53	4	488,15	2305,41	990,58	4596,24	119,29	3145,77	0	3718,4	0	2645,42	0	38670,79
10	1		22	3	0	0	0	0	3	668,42	366,78	102,93	161,23	30,94	0	0	1874,08	0	65,74	18,44	3288,56
11	1		22	2	2325,04	0	0	731,48	4	1335,95	789,49	299,85	883,2	65,44	3577,4	0	3569,67	0	1126,53	0	14704,05
12	1		22	1	2259	81,51	0	2022,81	4	201,39	1668,52	2112,65	2034,12	104,71	2504,42	0	3763,32	0	1376,11	885,28	19013,84
13	1		22	2	130,59	258,5	0	10944,3	4	365,3	1271,99	146,95	1407,04	204,36	2222,12	0	3420,33	0	1755,68	0	22127,16
14	1		23	3	386,42	0	0	302,23	4	972,73	986,42	1005,46	1251,76	20,53	0	453,65	1764,01	0	653,25	0	7796,46
15	3	26	23	2	1262,47	45,51	4	1356,47	5	942,69	2612,4	865,94	9506,02	406,64	938,23	1176,5	3125,24	0	5156,09	785,33	28179,53
16	2		23	2	35,1	0	0	3848,1	2	2480,62	2828,07	45,81	2475,56	355,78	594,94	0	492,02	2379,78	145,76	1238,67	16920,21
17	1		23	2	425,98	0	0	415,57	4	1133,37	1065,84	11,6	1433,22	243,93	3534,27	0	0	8924,17	731,19	0	17919,14
18	2	24	23	3	30,94	0	3	0	3	1629,55	982,85	458,4	1201,19	237,98	1399,61	0	5651,97	0	3701,45	0	15293,94
19	1		23	1	0	0	0	0	6	624,69	763,31	133,86	2369,07	51,76	0	25,29	3569,67	0	735,65	0	8273,3
20	2	23	23	2	0	446,51	4	47,6	4	262,67	2197,72	126,72	1853,55	456,62	2003,77	0	2677,25	0	1678,93	2586,22	14337,56
21	1		23	3	267,73	0	0	0	4	594,94	1188,1	181,16	500,35	412,89	0	0	2379,78	0	311,45	148,74	5985,14
22	1		23	2	1950,82	139,81	0	383,44	3	2007,05	2333,97	193,06	1428,46	183,84	174,02	556,27	5311,07	0	841,85	0	15503,66
23	1		23	2	0	0	0	38486,66	4	351,31	520,58	443,23	2828,07	85,37	594,94	0	6023,81	0	2408,63	837,09	52579,69
24	1		23	1	0	106,79	0	117,5	5	495,29	1714,33	40,16	1580,17	259,69	1786,62	74,37	2974,72	0	334,95	0	9484,59
25	2		23	2	2548,74	0	0	395,94	4	245,41	1035,8	0	3552,71	179,67	9406,96	0	0	2231,04	0	61,87	19658,14
26	1		23	2	337,04	98,46	0	1502,53	4	713,64	1866,94	330,49	3978,1	312,35	1131,88	0	3351,91	3272,19	3329,01	0	20224,54
27	1		24	2	0	0	0	148,44	4	504,22	252,55	65,44	3461,68	104,71	565,2	0	2974,72	0	144,27	0	8221,23
28	1		24	2	234,71	0	0	789,79	4	771,94	1332,68	132,97	1154,49	60,68	0	782,35	4658,41	0	2280,42	0	12198,44
29	1		24	3	184,43	59,49	0	1241,95	4	677,64	953,4	375,41	1140,21	470,01	0	220,13	2256,33	0	683,29	73,77	8336,06
30	2	27	24	2	475,96	0	3	8544,89	4	989,99	2031,44	697,57	3868,63	677,34	2840,56	267,73	4759,56	0	1706,6	502,43	27362,7
31	2	25	24	2	1664,36	58,9	3	3603,28	4	1293,11	3131,19	926,63	4004,57	357,86	4206,26	0	4352,02	0	1015,87	0	24614,05
32	1		24	2	0	47,6	0	0	1	630,34	1209,22	0	6309,39	253,74	0	26,77	2260,79	0	1342,79	1035,2	13115,84
33	1		24	1	653,25	144,87	0	29,75	6	638,67	839,76	0	3724,05	93,7	0	1011,41	3926,63	0	2546,06	361,73	13969,88
34	2	24	24	3	8,63	143,38	5	424,49	4	886,76	1960,34	1213,69	1357,66	198,71	2915,53	404,56	75,56	4581,07	1248,19	0	15418,57
35	2	25	24	2	3417,66	33,91	4	3393,27	5	526,23	3271,6	1048,89	3800,8	1141,4	3452,17	0	2911,96	5860,2	2350,33	121,37	31329,79
36	1		24	3	0	128,51	0	1192,57	4	451,86	2662,67	317,4	291,52	232,92	2424,4	0	0	2231,04	232,03	0	10164,92
37	2	23	24	3	237,68	218,64	5	3723,76	5	3638,68	2954,49	188	5875,08	321,57	1400,2	0	4908,29	0	555,38	0	24021,77
38	1		24	1	0	0	0	7043,25	5	471,79	555,38	156,47	3533,97	156,17	3734,76	0	3867,14	0	7961,55	0	27480,48
39	2		24	1	178,48	79,43	0	0	6	2399,71	2276,85	1960,94	9587,53	267,13	2998,52	0	5354,5	0	1860,39	0	26963,48
40	3		24	3	0	171,05	0	1687,56	4	1184,83	2815,87	899,85	4299,96	284,38	0	0	5305,71	0	2515,13	42,54	19206,88
41	1		24	1	0	995,64	0	1336,84	4	586,62	589,59	0	2273,28	59,49	0	72,88	2677,25	0	148,74	353,69	9094,02
42	2	27	24	1	1228,26	428,36	6	2728,12	5	1460,29	2447,3	1166,69	7426,39	293,31	1714,93	0	6246,92	0	2606,75	259,69	28007,01

Statistiques d'échantillons du Survey



Series: INCOME
échantillon 1
3816
Observations 3816

moyenne	23017.26
Median	19311.31
Maximum	293803.2
Minimum	0.000000
Std. Dev.	16412.63
Skewness	4.443018
Kurtosis	50.60523
Jarque-Bera	372889.9
probabilités	0.000000

Distribution de l'échantillon

Défnition

la distribution de θ_n est nomée distribution d'échantillon.

Examples

- Si les X_i suivent des tirages de Bernouilli(π) alors $\sum_{i=1}^{n} X_i$ (Le nombre de succès en n tirages) suit $\mathrm{Bin}(n, \pi)$.
- Si les X_i suivent $N(\mu, \sigma^2)$, $\overline{X} \sim N(\mu, \sigma^2/n)$.

En général la distribution *exacte* de l'échantillon est complexe. Elle

dépend de n de f_X , et du θ_n choisit.

Théorème de la moyenne d'échantillon

Théoreme

Si l'échantillon est aléatoire on a :

$$E[\overline{X}] = \mu, \qquad V(\overline{X}) = \frac{\sigma^2}{n}$$

Pour toute distribution de X avec mu et sigma comme moments.

Soit

$$E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_1] = n\mu$$
$$V\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} V(X_1) = n\sigma^2$$

Lorsque n croit, la variance de la moyenne d'échantillon décroit. Donc des échantillons plus grands permettent une estimation plus précise de la moyenne.

Sur la variance

Soit $Q^2 = n^{-1}\sum_{i=1}^{n}(X_i - \mu)^2$. qui n'est PAS une statistique

$$\mathrm{E}[Q^2] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2\right] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[(X_i-\mu)^2\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n}\mathrm{V}[X_i] = \frac{1}{n}\sum_{i=1}^{n}\sigma^2 = \sigma^2.$$

cependant

$$\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2\right] = \frac{n-1}{n}\sigma^2 \neq \sigma^2.$$

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

$$\mathrm{E}\left[S^2\right] = \frac{n}{n-1}E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2\right] = \sigma^2$$

Et on aura bien la variance d'échantillon centrée sur la vraie variance.

Attention : $E[S] = E[\sqrt{S^2}] \leq \sqrt{E[S^2]} = \sigma.$ Selon l'inégalité de Jensen, S est donc un estimateur biaisé qui tend à sous estimer σ.

Populations Normales

Theorem

si $X_1, \cdots, X_n$ est un échantillon d'une population normale de moyenne μ et variance σ^2, alors $\bar{X}$et S^2 sont des R.v. , avec $\bar{X}$normal de moyenne μ et variance σ^2/n et $(n-1)S^2/\sigma^2$ sera chi-deux avec $n-1$ degrés de liberté

- $\bar{X}$et

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

Ce théorème ne nous donne pas seulement les distributions de $\bar{X}$et S^2 pour une population normale Mais établit également le fait que les deux sont indépendants.

Dans les faits, cette indépendance de $\bar{X}$ et S^2 est une propriété unique de la distribution normale.

Inference et populations normales

Soit $X \sim N(\mu, \sigma)$. μ, σ inconnus mais l'on dispose d'un échantillon aléatoire sur X.
Que peut on dire de μ, σ^2? On sait que :

$$\overline{X} \sim \mathcal{N}(\mu, \sigma^2/n) \Longrightarrow Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1).$$

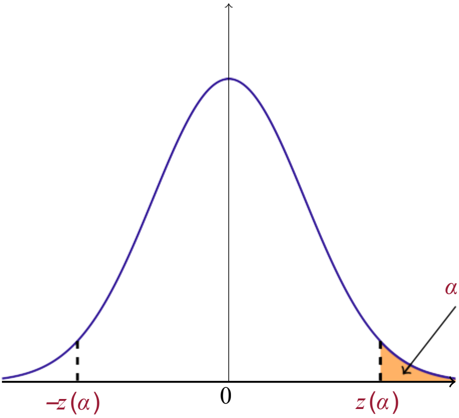
pour tout $\alpha \in (0, 1)$, supposons $z(\alpha)$ défini par

$$P(Z > z(\alpha)) = 1 - F(z(\alpha)) = 1 - \Phi(z(\alpha)) = \Phi(-z(\alpha)) = \alpha,$$

Donc $z(\alpha)$ est le $(1 - \alpha)$ième quantile de la loi standard normal.

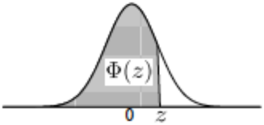
Ses valeurs sont tabulées $z(\alpha)$.

Figure: Distribution Normale



La table $z(\alpha)$ Entries

The c.d.f. of the standard normal distribution



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891

Inférence par intervalles de confiance [CI]

Concept des intervalles

- Règle basée sur les informations d'échantillon,
- Fournit un intervalle qui devrait contenir les valeurs vraies de l'estimateur,
- Avec une probabilité élevée.

pour un paramètre θ, avec une valeur $1-\alpha$ entre 0 et 1, le niveau de confiance est un estimator par intervalle est définir par deux r.v. $\hat{\theta}_A$ et $\hat{\theta}_B$ tel que

$$P(\hat{\theta}_A \leq \theta \leq \hat{\theta}_B) = 1-\alpha$$

pour deux valeurs concrètes de ces random variables, a et $b$on obtient un intervalle $[a; b]$ que l'on appelle intervalle de confiance à $100(1-\alpha)\%$ pour θ.

Inférence par intervalles de confiance [CI]

<u>Commentaires</u>

L'intervalle de confiance est associé à une probabilité , le niveau de confiance. Ses valeurs sont :

$$P(\hat{\theta}_A \leq \theta \leq \hat{\theta}_B) = 1 - \alpha$$

On obtient ces valeurs des quantiles de la distribution de notre estimateur, $\hat{\theta}$.

Pour calculer ces quantiles on doit connaitre la distribution de la relation entre θ et $\hat{\theta}$,

Cette distribution est la base de notre méthode par intervalles, elle dépend de :

- Le paramètre recherché (moyenne, variance),
- La distribution de la population,
- L'information disponible (par exemple d'autres paramètres).

Inférence sur la moyenne, σ^2 connue

Considérons un exemple simple

Hypothèses :

- On a un échantillon de n observations
- La population est normale
- On connait sa variance σ^2

objectif : construire un CI pour la moyenne de notre population

- pour un niveau $1 - \alpha$, la première étape est de trouver la distribution qui relie μ et X, ici :

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim \mathcal{N}(0, 1).$$

Inférence sur la moyenne, σ^2 connue

On construit donc un intervalle sur la distribution normale standard pour le niveau $z(\alpha/2)$ tel que :

$$P\left(\left|\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right| > z(\alpha/2)\right) = \alpha$$

ou,

$$P\left(-z(\alpha/2) \leq \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \leq z(\alpha/2)\right) = 1-\alpha.$$

soit,
$$P\left(\overline{X} - \frac{\sigma}{\sqrt{n}} z(\alpha/2) \leq \mu \leq \overline{X} + \frac{\sigma}{\sqrt{n}} z(\alpha/2)\right) = 1-\alpha$$

Et donc
$$\left[\overline{X} - \frac{\sigma}{\sqrt{n}} z(\alpha/2), \overline{X} + \frac{\sigma}{\sqrt{n}} z(\alpha/2)\right]$$

Constitues un $(1-\alpha)$ intervalle de confiance [CI] pour μ. Note: $z(\alpha)$ est décroissant en α. si σ^2 est connu, on peut le calculer.

Exemple

La température à laquelle le lait gèle est en moyenne $\mu = -.545^{\circ}C$, avec $\sigma = .008^{\circ}C$.
si le lait est coupé avec de l'eau, la distribution se décale vers la droite

Supposez qu'on ait 5 bouteilles.

En testant on obtient $\underline{x} = -.535^{\circ}C$.

Alors :

$$\mathrm{CI}_{.95} = \left[-.542, -.528\right], \qquad \mathrm{CI}_{.99} = \left[-.544, -.526\right]$$

$(z(.01/2) \approx 2.576;\ z(.05/2) \approx 1.96)$

Inference sur la moyenne, σ^2 inconnue

Si on ne connait pas σ^2 une idée intuitive serait de prendre S^2 à la place.

alors

$$\frac{\overline{X}-\mu}{S/\sqrt{n}} = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \Big/ \sqrt{\frac{S^2}{\sigma^2}}$$

N'est plus exactement normal,
Cependant on peut calculer que :

$$(n-1)\frac{S^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{\sigma}\right)^2 = \sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 - \left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right)^2$$

Une somme normale. Donc :

$$(n-1)\frac{S^2}{\sigma^2} \sim \chi^2(n-1)$$

Inférence sur la moyenne, σ^2 inconnue

En général, si $A \sim N(0, 1)$ et $B \sim \chi^2(p)$, alors

$$A \Big/ \sqrt{\frac{B}{p}} \sim t(p)$$

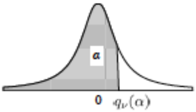
pour $t(p)$ une distribution t de student avec p degrés de liberté.

Dans notre cas,
$$\frac{\overline{X} - \mu}{S/\sqrt{n}} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \Big/ \sqrt{\frac{S^2}{\sigma^2}} \sim t(n-1).$$

la *distribution t* est symétrique et unimodale . pour $p = 1$, c'est une distribution de Cauchy, quand $p \to \infty$ cela converge vers une distribution normale.

La table $t(a; p)$

The quantile function of the Student's t distribution



ν \ α	0.600	0.700	0.750	0.800	0.850	0.900	0.925	0.950	0.975	0.990	0.995	0.999	0.9995
1	0.325	0.727	1.000	1.376	1.963	3.078	4.165	6.314	12.71	31.82	63.66	318.3	636.6
2	0.289	0.617	0.816	1.061	1.386	1.886	2.282	2.920	4.303	6.965	9.925	22.33	31.60
3	0.277	0.584	0.765	0.978	1.250	1.638	1.924	2.353	3.182	4.541	5.841	10.21	12.92
4	0.271	0.569	0.741	0.941	1.190	1.533	1.778	2.132	2.776	3.747	4.604	7.173	8.610
5	0.267	0.559	0.727	0.920	1.156	1.476	1.699	2.015	2.571	3.365	4.032	5.893	6.869
6	0.265	0.553	0.718	0.906	1.134	1.440	1.650	1.943	2.447	3.143	3.707	5.208	5.959
7	0.263	0.549	0.711	0.896	1.119	1.415	1.617	1.895	2.365	2.998	3.499	4.785	5.408
8	0.262	0.546	0.706	0.889	1.108	1.397	1.592	1.860	2.306	2.896	3.355	4.501	5.041
9	0.261	0.543	0.703	0.883	1.100	1.383	1.574	1.833	2.262	2.821	3.250	4.297	4.781
10	0.260	0.542	0.700	0.879	1.093	1.372	1.559	1.812	2.228	2.764	3.169	4.144	4.587
11	0.260	0.540	0.697	0.876	1.088	1.363	1.548	1.796	2.201	2.718	3.106	4.025	4.437
12	0.259	0.539	0.695	0.873	1.083	1.356	1.538	1.782	2.179	2.681	3.055	3.930	4.318
13	0.259	0.538	0.694	0.870	1.079	1.350	1.530	1.771	2.160	2.650	3.012	3.852	4.221
14	0.258	0.537	0.692	0.868	1.076	1.345	1.523	1.761	2.145	2.624	2.977	3.787	4.140
15	0.258	0.536	0.691	0.866	1.074	1.341	1.517	1.753	2.131	2.602	2.947	3.733	4.073
16	0.258	0.535	0.690	0.865	1.071	1.337	1.512	1.746	2.120	2.583	2.921	3.686	4.015
17	0.257	0.534	0.689	0.863	1.069	1.333	1.508	1.740	2.110	2.567	2.898	3.646	3.965
18	0.257	0.534	0.688	0.862	1.067	1.330	1.504	1.734	2.101	2.552	2.878	3.610	3.922
19	0.257	0.533	0.688	0.861	1.066	1.328	1.500	1.729	2.093	2.539	2.861	3.579	3.883
20	0.257	0.533	0.687	0.860	1.064	1.325	1.497	1.725	2.086	2.528	2.845	3.552	3.850

Inférence sur la moyenne, σ^2 inconnue

Soit $t(\alpha; p)$ le $(1-\alpha)$ième quantile de $t(p)$. alors

$$P\left(\overline{X} - \frac{S}{\sqrt{n}} t(\alpha/2; n-1) \leq \mu \leq \overline{X} + \frac{S}{\sqrt{n}} t(\alpha/2; n-1)\right) = 1 - \alpha$$

$$\left[\overline{X} - \frac{S}{\sqrt{n}} t(\alpha/2; n-1), \overline{X} + \frac{S}{\sqrt{n}} t(\alpha/2; n-1)\right]$$

Constituera un intervalle de confiance $(1-\alpha)$ pour μ.
À cause des queues de distributions plus larges, l'estimateur est moins précis qu'avec σ *connu*.
Leur différence converge vers 0 quand $n \rightarrow \infty$. Example du

lait avec S =0,008 :

$$t(.05/2; 5-1) \approx 2.776$$

$$\mathrm{CI}_{.95} = \left[-.535 - 2.776 \frac{.008}{\sqrt{5}}, -.535 + 2.776 \frac{.008}{\sqrt{5}}\right] = [-.545, -.525]$$

Inference en Populations Non-Normales

En général, l'inférence sur un échantillon finie est complexe.

Définir la distribution exacte d'un échantillon fini X s'avère même souvent impossible.

Une alternative consiste à approximer la distribution de X pour l'inférer. L'approche classique repose sur la théorie asymptotique (large-n) .

On considère le comportement de la séquence stochastique $\{\theta_n\}$ as $n \rightarrow \infty$.

Des propriétés très larges peuvent être démontrées :

Sequences de r.v.'s

Les sequences sont des décomptes d'objets (number, r.v., etc.) arrangés en ensembles ordonnés.

On rencontre souvent des sequences de r.v.'s et est interessé par leurs comportements, notamment en termes de convergences.

Une r.v. est une fonction de l'échantillon S vers R, une sequences de r.v.'s est juste une sequence de fonctions.

Considerons la sequence de r.v.'s

$$\{X_n\} = \{X_n,\ n \geq 1\} = \{X_1, X_2, \cdots, X_n\}$$

Chaque X_i a une pdf $f_i(x)$ et a une cdf $F_i(x)$.

On s'intéresse à comment $\{X_n\}$ converge when $n \to \infty$
Vers une r.v. X avec pour pdf $f_X(x)$ et cdf $F_X(x)$.

Convergence de Séquences de nombres réels

Toute fonction $a : \mathbb{N} \to \mathbb{R}$.

Comment ces séquences se comportent quand n devient large ?

définition: Convergence de sequence de nombres réels

La séquence $\{s_n\}$ converge vers s, si pour tout $g > 0$ il existe un entier positif N tel que

$$|s_n - s| < g, \qquad \forall n \geq N$$

s est appelé limite de la séquence

Convergence en probabilité

Soit $\{X_n\}$ une séquence de r.v. (e.g., un échantillon statistique). La première limite est celle de convergence en probabilité.

Définition

$\{X_n\}$ converges en probabilités ($X_n \xrightarrow{p} X$) si, pour tout $g > 0$,

$$\lim_{n \to \infty} P(|X_n - X| \geq g) = 0, \qquad \text{i.e.,} \qquad \lim_{n \to \infty} P(|X_n - X| < g) = 1$$

La probabilité que X_n soit plus loin que g de X converge vers zero quand n devient grand.

X est appellé probabilité limite de la sequence. Soit :

$$\operatorname*{plim}_{n \to \infty} X_n = X.$$

Convergence quasi certaine

La Convergence quasi certaine ressemble beaucoup à la précédente, mais ses hypothèses sont plus fortes.

Definition

On dit que $\{X_n\}$ converges de manière quasi certaine ($X_n \xrightarrow{a.s} X$) si,

$$P\left(\lim_{n \to \infty} X_n = X\right) = 1.$$

Exemple

Le nombre de jour dans une année non bissextile converges vers 365

$$X_n \xrightarrow{a.s.} X$$

Convergence en Distribution

Convergence en distribution apparait beaucoup dans les modèles économétriques de par l'utilisation du théorème central limite

Definition

une sequence de r.v.'s $\{X_n\}$ converge en distribution vers la r.v. X si $n \to \infty$ $X_n \xrightarrow{d} X$ si ,

$$\lim_{n \to \infty} F_{X_n}(x) = F_X(x)$$

Pour tout point continue de F_X .

Cela signifie que la probabilité que X appartienne à un certain intervalle est très proche de la probabilité que Xn soit dans cet intervalle pour n suffisamment grand.

Propriétés de Convergences

Sans démonstrations retenez
que :

$$X_n \xrightarrow{a.s.} X \implies X_n \xrightarrow{p} X \implies X_n \xrightarrow{d} X$$

La convergence en distribution est la plus faible

Loi des grands nombres (faible) [LLN]

Théorème (Khintchine) — Soit (X_n) une suite de variables aléatoires réelles indépendantes et de même loi admettant une espérance μ. La moyenne empirique $\overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k$ converge en probabilité vers l'espérance : pour tout $\varepsilon > 0$, on a $\lim_{n \to +\infty} \mathbf{P}\left(\left|\overline{X}_n - \mu\right| > \varepsilon\right) = 0$.

Ce résultat assure en particulier que la moyenne empirique est un estimateur convergent de l'espérance.

si μ_X existe.

pour tout $g > 0$,

$$\lim_{n \to \infty} P(|X_n - \mu| \geq g) = 0, \qquad \text{i.e.,} \qquad \lim_{n \to \infty} P(|X_n - \mu| < g) = 1$$

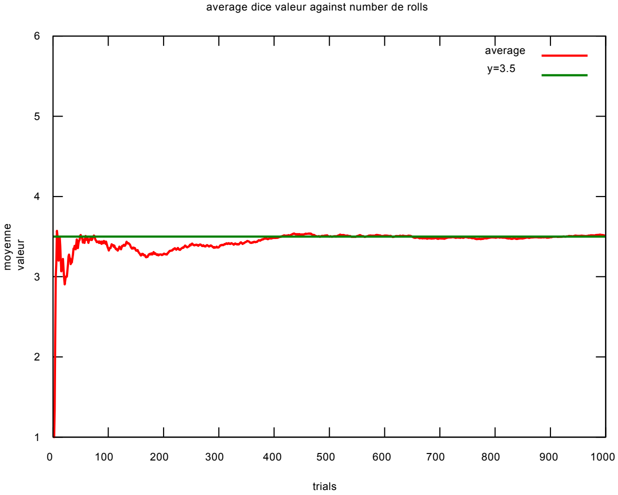
Loi des grands nomres forte [LLN]

La loi forte des grands nombres postule que la moyenne d'échantillon converge presque certainement vers l'espérance

$$X_n \xrightarrow{a.s.} \mu \quad \text{quand } n \to \infty$$

Théorème — $\mathbb{P}\left(\lim_{n\to+\infty} \overline{X}_n = E(X)\right) = 1.$

LLN: échantillon moyenne de lancer de dés



Central Limit Theorem [CLT]

Le théorème central établit la convergence en loi de la somme d'une suite de variables aléatoires vers la loi normale. Intuitivement, une somme de variables aléatoires identiques et indépendantes tend vers une variable aléatoire Normale. Pour N assez grand

Afin de formuler mathématiquement cette approximation, nous allons poser

$$\overline{X}_n = \frac{S_n}{n} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

et

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}},$$

Donc par le CLT

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \leq z\right) = \Phi(z).$$

[CLT]: Exemple

Soit X_i'des tirages de $Bernoulli(p)$.

alors $E[X_i] = p$ et $\mathrm{Var}(X_i) = p(1-p)$.

et, $Y_n = X_1 + X_2 + \ldots + X_n$ suit $Bin(n, p)$. donc,

$$Z_n = \frac{Y_n - np}{\sqrt{np(1-p)}},$$

où $Y_n \sim Bin(n, p)$.

On voit en simulant Yn que cela tends vers une forme de courbe normale quand N augmente

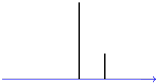
Assumptions:

- $X_1, X_2 \ldots$ are iid Bernoulli(p).
- $Z_n = \dfrac{X_1 + X_2 + \ldots + X_n - np}{\sqrt{np(1-p)}}$.

We choose $p = \frac{1}{3}$.

$$Z_1 = \frac{X_1 - p}{\sqrt{p(1-p)}}$$

PMF of Z_1

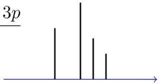


$$Z_2 = \frac{X_1 + X_2 - 2p}{\sqrt{2p(1-p)}}$$

PMF of Z_2

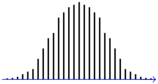
$$Z_3 = \frac{X_1 + X_2 + X_3 - 3p}{\sqrt{3p(1-p)}}$$

PMF of Z_3



$$Z_{30} = \frac{\sum_{i=1}^{30} X_i - 30p}{\sqrt{30p(1-p)}}$$

PMF of Z_{30}



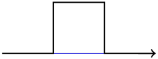
[CLT]: Exemple continu

Assumptions:

- $X_1, X_2 \dots$ are iid Uniform(0,1).
- $Z_n = \dfrac{X_1 + X_2 + \ldots + X_n - \frac{n}{2}}{\sqrt{\frac{n}{12}}}$.

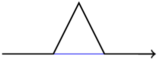
$$Z_1 = \frac{X_1 - \frac{1}{2}}{\sqrt{\frac{1}{12}}}$$

PDF of Z_1



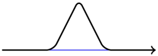
$$Z_2 = \frac{X_1 + X_2 - 1}{\sqrt{\frac{2}{12}}}$$

PDF of Z_2



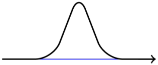
$$Z_3 = \frac{X_1 + X_2 + X_3 - \frac{2}{3}}{\sqrt{\frac{3}{12}}}$$

PDF of Z_3

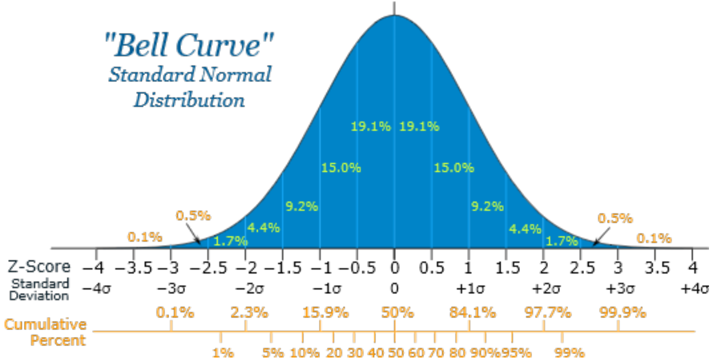


$$Z_{30} = \frac{\sum_{i=1}^{30} X_i - \frac{30}{2}}{\sqrt{\frac{30}{12}}}$$

PDF of Z_{30}



Annexe


"Bell Curve"
Standard Normal Distribution
19.1% 19.1%
15.0% 15.0%
9.2% 9.2%
4.4% 4.4%
1.7% 1.7%
0.5% 0.5%
0.1% 0.1%
Z-Score −4 −3.5 −3 −2.5 −2 −1.5 −1 −0.5 0 0.5 1 1.5 2 2.5 3 3.5 4
Standard Deviation −4σ −3σ −2σ −1σ 0 +1σ +2σ +3σ +4σ
Cumulative Percent 0.1% 2.3% 15.9% 50% 84.1% 97.7% 99.9%
1% 5% 10% 20 30 40 50 60 70 80 90% 95% 99%