Eléments de statistiques pour les data sciences

Cours 5 : la distribution de Student

Dr. Ph. Müllhaupt

IGM - EPFL

—

Plan

loi normale $\mathcal{N}(\mu, \sigma)$

Soit une variable aléatoire

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

sa densité de probabilité est donnée (rappel) par

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

et introduisons un changement de variable aléatoire pour obtenir une variable centrée et réduite $\mathcal{N}(0, 1)$:

$$Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1) \qquad g(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

On peut alors utiliser la table de la loi normale.

1 échantillon de taille 1 de la loi normale $\mathcal{N}(0,1)$

Soit une valeur x (échantillon de dimension 1) d'une variable aléatoire $X \sim \mathcal{N}(0,1)$.

probabilité de tomber dans un intervalle, valeur 95 %

Soit $z > 0$,

$$0.95 = P(|Z| < z) = 1 - 2 \cdot P(Z > z) = 1 - 2 \cdot 0.025 = 2 \cdot 0.475 = 2 \cdot P(0 < Z \leq z)$$

définit un intervalle qui garantit que si l'expérience est répétée un grand nombre de fois N, il y aura asymptotiquement $0.95 \times N$ échantillons qui tombent dans l'intervalle.

Pour déterminer l'intervalle $[-z; +z]$, on utilise la table pour déterminer z tel que

$$0.475 = P(0 < Z \leq z)$$

longueur de l'intervalle

garantissant 95 % de probabilité de tomber dans l'intervalle

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07
0.	0.	0.004	0.008	0.012	0.016	0.002	0.0239	0.0279
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596		
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.475	0.4756

intervalle demandé pour $\mathcal{N}(0.1)$: $\boxed{[-1.96; 1.96]}$

changement de référentiel

Jusqu'à présent on observe l'échantillon du point de vue de la distribution fixe $\mathcal{N}(0,1)$. La moyenne $\mu = 0$ est l'origine du référentiel fixe.

$$\mathrm{O} = \mu \,\hat{x} \qquad \mu = 0$$

Changeons de référentiel, et plaçons une nouvelle origine centré sur l'échantillon

$$\mathrm{O}' = x_k \,\hat{x} \qquad \mu = 0$$

On se rend compte que lorsque on centre l'origine chaque fois sur la réalisation de la variable aléatoire, c'est la moyenne fixe qui semble devenir une variable aléatoire !

$$\text{intervalle de confiance à 95 \% :} \qquad \mu \in [x_k - 1.96\,;\, x_k + 1.96]$$

1 échantillon de taille 1 – intervalle de confiance à 95 %

lorsque $X \sim \mathcal{N}(\mu, \sigma^2)$

$$Z = \frac{X - \mu}{\sigma} \qquad Z \sim \mathcal{N}(0,1)$$

$$[z_k + 1.96\,;\; z_k - 1.96]$$

$$\left[\frac{x_k - \mu}{\sigma} - 1.96; \frac{x_k - \mu}{\sigma} + 1.96\right]$$

$$\boxed{x_k \in [\mu - 1.96\,\sigma; \mu + 1.96\,\sigma]}$$

changement de référentiel :

$$\boxed{\mu \in [x_k - 1.96\,\sigma\,;\; x_k + 1.96\,\sigma]}$$

stabilité de la loi normale

Si on a deux variables indépendantes X_1 et X_2 distribuées selon

$$X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2) \qquad X_2 \sim \mathcal{N}(\mu_1, \sigma_2^2)$$

alors

$$X_1 + X_2 \sim \mathcal{N}\left(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2\right)$$

stabilité de la loi normale

démonstration par la convolution dans le cas $X \sim \mathcal{N}(0,1)$

$$
\begin{aligned}
h(x) &= \int_{\xi=-\infty}^{+\infty} f(\xi)\, g(x-\xi)\, d\xi \\
&= \frac{1}{2\pi} \lim_{a\to\infty} \int_{-a}^{+a} e^{-\frac{\xi^2}{2}} e^{-\frac{(\xi-x)^2}{2}}\, d\xi \\
&= \frac{1}{2\pi} e^{-x^2/4} \sqrt{\pi} \lim_{a\to\infty} \left(\int -\frac{1}{\sqrt{\pi}} e^{-\left(a-\frac{x}{2}\right)^2} dx + \int \frac{1}{\sqrt{\pi}} e^{-\left(a+\frac{x}{2}\right)^2} dx \right) \\
\mathrm{Erf}(x) &= \int \frac{2}{\sqrt{\pi}} e^{-x^2}\, dx \\
h(u) &= \frac{1}{4\pi} e^{-\frac{x^2}{4}} \sqrt{\pi} \lim_{a\to\infty} \left(\mathrm{Erf}(a - x/2) + \mathrm{Erf}(a + x/2)\right) = \frac{1}{2\pi} e^{-\frac{x^2}{4}} \sqrt{\pi} \\
&= \frac{1}{\sqrt{2\pi \cdot 2}} e^{-\frac{x^2}{2\cdot 2}} \qquad \Rightarrow \qquad X + X \sim \mathcal{N}(0,2)
\end{aligned}
$$

stabilité de la loi normale

démonstration en utilisant la fct. génératrice des moments

définition de la fonction génératrice des moments

$$M_X(u) \triangleq \mathbb{E}[e^{uX}] = \int_{-\infty}^{+\infty} e^{ux} f(x) dx$$

c'est la transformée de Laplace bilatérale avec changement de signe $s = -u$ et variable $u \in \mathbb{R}$ au lieu de $s \in \mathbb{C}$ pour la transf. de Laplace.

Pour la loi normale

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \leftrightarrow e^{u\,\mu + \sigma^2 \frac{u^2}{2}}$$

moyenne d'un échantillon unique de taille n

Soit un échantillon unique de n valeurs d'une variable aléatoire X :

$$x_1, x_2, \ldots, x_n$$

définition

la moyenne de l'échantillon est donnée par

$$\boxed{\bar{x} \triangleq \frac{1}{n} \sum_{k=1}^{n} x_k}$$

1 éch. de taille n – intervalle de confiance à 95 %

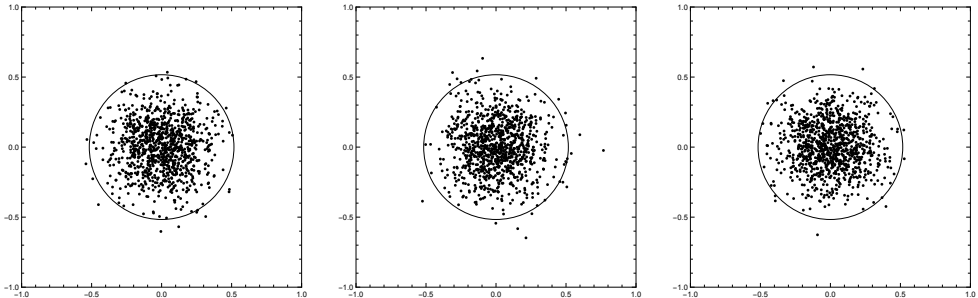
statistique $\bar{X}$

$$\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}$$

$$\boxed{\mu \in \left[\bar{x} - 1.96\,\frac{\sigma}{\sqrt{n}}\,;\,\bar{x} + 1.96\,\frac{\sigma}{\sqrt{n}}\right]}$$

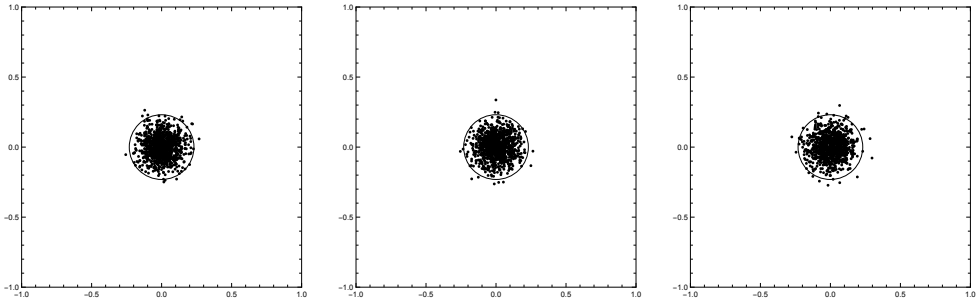
effet de la taille de l'échantillon n sur la précision

i.e. réduction de la variance



effet de la taille de l'échantillon n sur la précision

i.e. réduction de la variance



précision et nombre d'échantillons

Soit L la précision requise. On aimerait

$$\mu \in [\bar{x} - L\,;\, \bar{x} + L]$$

ainsi

$$L = 1.96\frac{\sigma}{\sqrt{n}} \approx \frac{2\sigma}{\sqrt{n}}$$

ce qui donne

$$n \approx \frac{4\sigma^2}{L^2}$$

Il faut connaître la variance σ^2 et être sûr que les données suivent une distribution normale $\mathcal{N}(\mu, \sigma^2)$. On ne connaît pas μ mais on l'estime avec $\bar{x}$ une réalisation de la variable aléatoire $\bar{X}$.

variance de l'échantillon

définition

la variance de l'échantillon est définie par

$$s^2 \triangleq \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x})^2$$

REMARQUE : on divise par $n-1$ et non par n

formule utile pour la somme de carrés

$$\sum_{k=1}^{n}(x_k - \bar{x})^2 = \sum_{k=1}^{n} x_k^2 - \bar{x}\sum_{k=1}^{n} x_k$$

démonstration

exercice

la statistique t

on ne connaît plus la variance σ^2

objectif

Il s'agit de quantifier la signification de l'écart de la moyenne de l'échantillon par rapport à 0.

définition

$$t \triangleq \frac{\bar{x}}{\sqrt{\frac{s^2}{n}}}$$

la distribution de Student

Si X est variable aléatoire distribuée selon une loi normale $\mathcal{N}(0, \sigma)$ avec une variance σ connue exactement, alors la probabilité que X/σ excède n'importe quelle valeur peut être calculée à partir d'une loi normale $\mathcal{N}(0, 1)$.

Mais si σ n'est pas connu, mais estimé en utilisant s, alors la distribution X/s n'est plus normale (en particulier lorsque n est petit). La vraie valeur a été divisée par s/σ, qui introduit une erreur.

En utilisant les fonctions génératrice des moments, il est relativement aisé de démontrer que s^2/σ^2 est distribuée selon la loi du $\chi^2_{(n)}/n$ à n degrés de liberté.

La distribution de X/s est donc calculable et dépend des degrés de liberté nécessaires pour l'estimation de σ.

hypothèse concernant le modèle probabiliste

... et un peu d'histoire, parenté

hypothèse

Les échantillons x_i proviennent d'une distribution normale $\mathcal{N}(\mu, \sigma^2)$ dont on ne connait ni la vraie moyenne μ ni la vraie variance σ^2.

Gauss a montré que la distribution de l'estimateur de la moyenne $\bar{x}$ obéit a une loi normale centrée sur la vraie moyenne μ mais de variance divisée par n.

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Helmert en 1875 a montré que la distribution de la variance de $\bar{X}$ était indépendante de μ. On peut montrer que la variable

$$u = \frac{1}{2}(n-1)\frac{s^2}{\sigma^2}$$

suit une distribution du χ^2 avec $n-1$ degrés de liberté (comme on a pas nécessairement $\mu = 0$, on perd un degré de liberté).

densité de probabilité

William S. Gossett a démontré que la statistique

$$t = \frac{\bar{x}}{\sqrt{\frac{s^2}{n}}}$$

sous hypothèse que les échantillons soient distribués selon $X \sim \mathcal{N}(\mu, \sigma^2)$ obéit avec $\nu = n - 1$ à la

fonction de densité de probabilité pour t

$$\boxed{f(t) = \frac{1}{\sqrt{\pi\,\nu}} \frac{\frac{\nu-1}{2}!}{\frac{\nu-2}{2}!} \frac{1}{\left(1 + \frac{t^2}{\nu}\right)^{(\nu+1)/2}}}$$

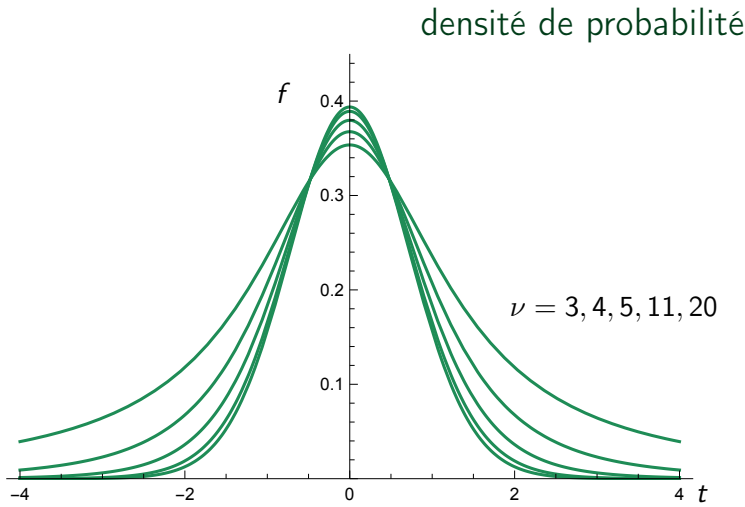


Figure – Graphique de la distribution $f(t)$ de la loi de Student pour différentes valeurs de l'indice ν correspondant au degré de liberté, en fonction de la variable t. Le graphique s'approche d'une gaussienne lorsque $\nu \to \infty$. Les queues sont épaissent lorsque ν est petit

table de la loi de Student

	probabilité qu'une valeur supérieure soit atteinte, sans compter le signe								
ν	0.500	0.400	0.200	0.100	0.050	0.025	0.010	0.005	0.001
1	1.	1.376	3.078	6.314	12.71	25.45	63.66	127.3	636.6
2	0.8165	1.061	1.886	2.92	4.303	6.205	9.925	14.09	31.6
3	0.7649	0.9785	1.638	2.353	3.182	4.177	5.841	7.453	12.92
4	0.7407	0.941	1.533	2.132	2.776	3.495	4.604	5.598	8.61
5	0.7267	0.9195	1.476	2.015	2.571	3.163	4.032	4.773	6.869

Table – Table de la loi de Student, valeurs de t, (test à deux queues)

paramètre α de la table de la loi de Student

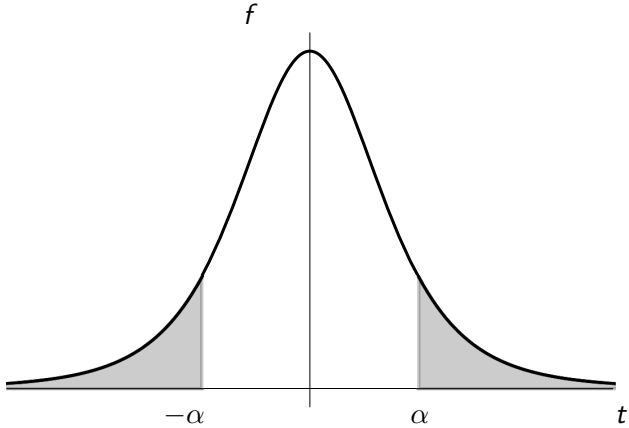


Figure – la table donne la valeur $t = \alpha$ de l'abscisse pour une valeur de la probabilité donnée en première ligne de la table. Cette probabilité correspond à la surface hachurée. L'indice ν de la table (pas représenté dans la figure) indique les degrés de liberté.

probabilité P_α de la table de la loi de Student

La table indique P_α pour une valeur α se situant en première ligne. Cette valeur est donnée par

$$P_\alpha = P(X \leq -\alpha) + P(X \geq \alpha) = \int_{-\infty}^{-\alpha} f(t)\,dt + \int_{\alpha}^{+\infty} f(t)\,dt$$

et correspond à la surface de la zone grise de la figure.

exemple d'application

Soit $n = 5$ échantillons d'une valeur numérique entière. Il s'agit de déterminer si ces données sont cohérentes avec une distribution gaussienne de variance inconnue mais de moyenne nulle. On a indiqué l'indice de l'échantillon en première ligne.

k	1	2	3	4	5
x_k	-137	63	-84	-95	110

exemple d'application

moyenne d'échantillon

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{5} x_k = \frac{-137+63-84-95+110}{5} = -\frac{-363}{5} = -72.6$$

variance d'échantillon

$$s^2 = \frac{1}{n-1}(x_k - \bar{x})^2 = 6141.3$$

Pour calculer s^2, on commence par calculer $(n-1)s^2 = \sum_k x_k^2 - \bar{x}\sum_k x_k$:

$$\begin{aligned} 4s^2 &= 137^2 + 63^2 + 84^2 + 95^2 + 110^2 - \frac{-363.5}{5} \times (-363) = \sum X^2 - \bar{x}\sum X \\ &= 50919 - \frac{-363}{5} \cdot (-363) = 24565.2 \end{aligned}$$

ce qui correspondond bien à 4×6141.3.

test de signification

Comme pour le χ^2 on peut prendre un seuil de 5 % pour caractériser ce qui donne de l'évidence pour une déviation grande de la moyenne nulle. Pour ν, comme on a utilisé les données pour calculer s, on retranche 1 au nombre de mesures.

$$\nu = 5 - 1 = 4$$

$$t = \frac{\bar{x}}{\sqrt{\frac{s^2}{n}}} = \frac{-72.6}{\sqrt{\frac{6141.3}{5}}} = -2.07$$

Et en lisant dans la table le long de la ligne $\nu = 4$ de la table de Student, on constate

$$\text{NS} = P\{|t_{(4)}| \geq 2.07\} \qquad 0.20 < \text{NS} < 0.10$$

Les données ne montrent pas d'évidence significative contre l'hypothèse de la moyenne nulle.

test de l'écart entre deux séries

Le test précédent s'applique naturellement à deux séries de variables. On aimerait tester si les deux séries s'écartent significativement l'une de l'autre.

exemple

Dans 8 usines, le nombre d'heures moyennes perdues sur une année suite à des accidents a été reporté. Le nombre A correspond à celles après la prise de mesure de protection et B celle sans les mesures de protection. On demande de comparer.

usine	k	1	2	3	4	5	6	7	8
avant	B_k	48.5	79.2	25.3	19.7	130.9	57.6	88.8	62.1
après	A_k	28.7	62.2	28.9	0.0	93.5	49.6	86.3	40.2
différence	X_i	19.8	17.0	-3.6	19.7	37.4	8.0	2.5	21.9

comparaison entre deux séries

exemple

Il faut assumer que les $X_k \triangleq B_k - A_k$ sont des variables aléatoires indépendantes toutes distribuée selon $\mathcal{N}(\mu, \sigma^2)$. Pour cet exemple :

$$n = 8 \qquad \bar{x} = 15.34 \qquad s^2 = 164.12 \qquad H_0 : \mu = 0$$

$$t = \frac{15.34}{\sqrt{\frac{164.12}{8}}} = 3.39$$

$$\text{NS} = P\{|t_{(7)}| \geq 3.39\} \qquad 0.025 < \text{NS} < 0.01$$

Il y a de l'evidence que les mesures ont été efficace. En effet, il y a de l'evidence que l'hypothèse soit fausse car NS est sous les 5 %.

Résumé

différence entre variance connue σ^2 et estimée s^2

$$Z \triangleq \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim \mathcal{N}(0,1)$$

$$T \triangleq \frac{\bar{X} - \mu}{\sqrt{\frac{S^2}{n}}} \sim t_{(n-1)}$$

comparaison de deux moyennes

deux séries, deux échantillons de taille n et m

échantillon 1 : $X_1, X_2, \ldots, X_n \sim \mathcal{N}(\mu_1, \sigma_1^2)$

échantillon 2 : $Y_1, Y_2, \ldots, Y_m \sim \mathcal{N}(\mu_2, \sigma_2^2)$

quelques résultats ...

concernant les moyennes et les variances

4 statistiques qui sont des variables aléatoires :

$$\bar{X} \triangleq \frac{1}{n}\sum X_i \qquad V_1 \triangleq \sum (X_i - \bar{X})^2$$

$$\bar{Y} \triangleq \frac{1}{m}\sum Y_i \qquad V_2 \triangleq \sum (Y_i - \bar{Y})^2$$

dont les distributions sont :

$$\bar{X} \sim \mathcal{N}\left(\mu_1, \frac{\sigma_1^2}{n}\right) \qquad V_1/\sigma_1^2 \sim \chi^2_{(n-1)}$$

$$\bar{Y} \sim \mathcal{N}\left(\mu_2, \frac{\sigma_2^2}{m}\right) \qquad V_2/\sigma_2^2 \sim \chi^2_{(m-1)}$$

$$\phi \triangleq \mu_1 - \mu_2$$

$$\hat{\phi} \triangleq \bar{X}_1 - \bar{X}_2 \sim \mathcal{N}\left(\phi, \frac{1}{n}\sigma_1^2 + \frac{1}{m}\sigma_2^2\right)$$

$$H_0 : \phi = 0$$

$$\boxed{Z \triangleq \frac{\hat{\phi} - \phi}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \sim \mathcal{N}(0,1)}$$

Il faut connaître σ_1^2 et σ_2^2