

GC – Probabilités et Statistique

<http://moodle.epfl.ch/course/view.php?id=14271>

Cours 9

- Modélisation statistique
- Distribution t de Student, t -test

Modèles statistiques

- Un **modèle statistique** est une description mathématique approximative du mécanisme qui a généré les observations, qui tient compte des *erreurs aléatoires et imprévisibles* :
 - donne une représentation *idéalisée* de la réalité
 - fait des *suppositions explicites* (qui peuvent être **fausses** !!) sur les processus étudiés
 - permet un raisonnement *abstrait*
- Le modèle s'exprime par une *famille de distributions théorique* qui contient des cas 'idéaux' pour les VAs inclues
 - p. ex. : jets d'une pièce ...
- Un modèle utile offre un *bon compromis* entre
 - description *juste* de la réalité (paramètres nombreux, suppositions correctes)
 - *facilité* de manipulation mathématique
 - production de solutions/prévisions *proches* de l'observation(s)

Un modèle simple

Un cas simple : on effectue plusieurs mesures d'une quantité physique μ , p. ex. longueur d'un champ, taille d'une personne ...

- De telles mesures possèdent en général une composante *aléatoire* due aux *erreurs de mesure*

- Un mécanisme d'erreur possible :

$$\begin{array}{ccccccc} \text{mesure} & = & \text{vraie valeur théorique} & + & \text{erreur de mesure} \\ y & = & \mu & + & \epsilon \end{array}$$

- c.-à-d. : des mesures avec des *erreurs additives*
- S'il n'y a pas d'erreur *systématique* (biais), l'erreur aléatoire doit être 'centrée' ($E[\epsilon] = 0$)
- Souvent raisonnable de penser que *la précision* de chaque mesure est *la même* ($\text{Var}(\epsilon) = \sigma^2$ pour chaque mesure)
- *Une spécification possible* pour la distribution de l'erreur est la loi *normale* $N(0, \sigma^2)$
- **All models are wrong; some are useful**

Estimation des paramètres inconnus

- Une fois un modèle est choisi, l'intérêt se tourne vers l'estimation des inconnus : *les paramètres du modèle*
- On observe des *réalisations* d'une VA dont on connaît la distribution (sauf les valeurs des paramètres)
- Donc, on doit *estimer* les paramètres à l'aide des observations $X_1, \dots, X_n$
- $\hat{\mu} = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$
- $\hat{\sigma}^2 = S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$
- L'estimateur S^2 est *nonbiaisé* pour σ^2 , et est *indépendant* de celui de μ ($\bar{X}$)

Révision : Théorème Central Limite (TCL)

Théorème (TCL) : Soient $X_1, X_2, \dots$ des variables aléatoires indépendantes et identiquement distribuées (iid), et telles que $E[X_i] = \mu$ et $Var(X_i) = \sigma^2 < \infty$ existent. Alors, la distribution de

$$\frac{X_1 + \dots + X_n - n\mu}{\sigma\sqrt{n}}$$

se rapproche d'une distribution normale lorsque $n \rightarrow \infty$.

C.-à-d. : Plus n est grand ('suffisamment grand'), plus *la loi de la somme (ou la moyenne)* se rapproche d'une distribution normale.

$$\Rightarrow \text{Donc, } \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right); \quad \hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

À propos des échantillons petits...

- Le z-test que nous avons étudié suppose que la distribution d'échantillonnage de la statistique de test T est *normale*
 - soit exactement
 - soit approximativement, par le TCL
- Toutefois, si l'écart-type de la population σ est *inconnu* et la taille de l'échantillon est *faible* (moins de 30-40, p. ex.), alors la *vraie distribution d'échantillonnage* de T possède des *queues qui s'étendent plus loin* que la distribution normale
- Dans ce cas, on utilise le *t-test*

'Student' (= William Sealy Gosset)

W. S. Gosset



Guinness



Distribution de T quand σ^2 est inconnue

- Rappelons la statistique de test $T = (\bar{X} - \mu_H)/(\sigma/\sqrt{n})$
- **Si** la taille de l'échantillon n est 'suffisamment grande', alors sous H , $T \sim N(0,1)$ *quelle que soit la distribution de X* (TCL)
- **Si** les observations $X_1, \dots, X_n \sim N(\mu_H, \sigma^2)$, alors $T \sim N(0,1)$ pour σ^2 *connue*, *quelle que soit la taille de l'échantillon n*
- **MAIS :** Si la taille de l'échantillon n est *petite*, et la variance σ^2 est *inconnue*, la *vraie* distribution de T a *davantage de variabilité* que la distribution normale (due à l'estimation *imprécise* de σ basée sur peu d'obs)
- Dans le cas **(1)** $X_1, \dots, X_n \sim N(\mu_0, \sigma^2)$; **(2)** n est petite; et **(3)** σ^2 est inconnue : alors $T = \frac{\bar{X} - \mu_H}{s/\sqrt{n}} \sim t_{n-1}$, la distribution t de Student, avec $n-1$ *degrés de liberté* (df; 'degrees of freedom')
- (La distribution de T dépend du nombre d'observations n)

Distribution t de Student

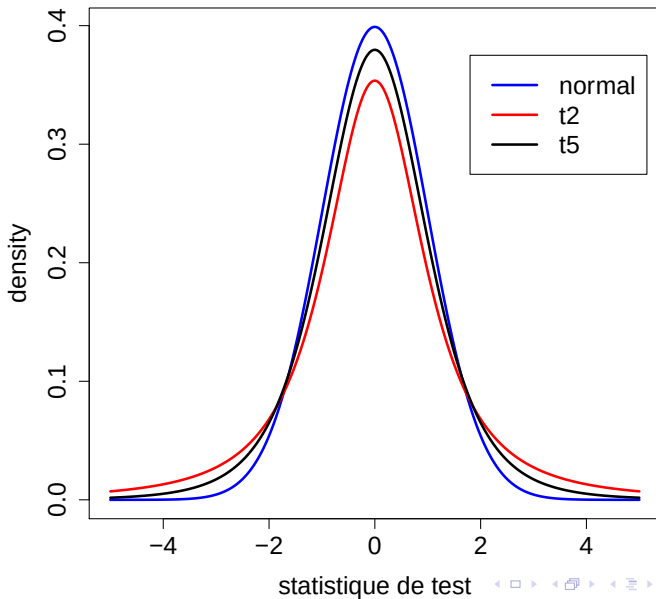


Table de la distribution t de Student

t Table

cum. prob one-tail two-tails	$t_{.50}$ 0.50	$t_{.75}$ 0.25	$t_{.80}$ 0.20	$t_{.85}$ 0.15	$t_{.90}$ 0.10	$t_{.95}$ 0.05	$t_{.975}$ 0.025	$t_{.99}$ 0.01	$t_{.995}$ 0.005	$t_{.999}$ 0.001	$t_{.9995}$ 0.0005
df	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
Z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

Intervalle de confiance

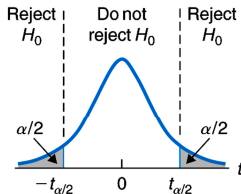
Dans le cas

- 1 $X_1, \dots, X_n \sim N(\mu, \sigma^2)$
 - 2 n est petite ; et
 - 3 σ^2 est inconnue :
- on peut faire un *intervalle de confiance (IC)* comme avant, mais **en utilisant la distribution t au lieu de la normale (z)**
 - IC pour la *moyenne* de la population : $\bar{x} \pm \boxed{t_{n-1, 1-\alpha/2}} \boxed{s} / \sqrt{n}$

Test d'hypothèses : trouver la région de rejet

$$H: \mu = \mu_H$$

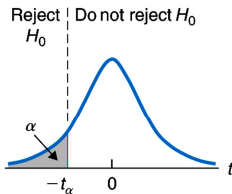
$$A: \mu \neq \mu_H$$



Two-tailed

$$H: \mu = \mu_H$$

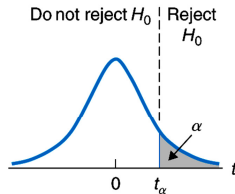
$$A: \mu < \mu_H$$



Left-tailed

$$H: \mu = \mu_H$$

$$A: \mu > \mu_H$$



Right-tailed

Exemple

Exemple 9.1

Prise quotidienne d'énergie (kJ) pour 11 femmes :

5260 5470 5640 6180 6390 6515 6805 7515 7515 8230 8770

- Faire un IC de 95% pour la moyenne prise (kJ) de la population des femmes ...
- Tester l'hypothèse que la moyenne est égale à la valeur recommandée (7725 kJ) ...

Test

1

2

3

4

5

Test de comparaison de 2 moyennes : variances égales

- On veut comparer les moyennes de deux suites de mesures :

- Groupe 1 (p. ex. 'contrôle') : $x_1, \dots, x_n$

- Groupe 2 (p. ex. 'traitement') : $y_1, \dots, y_m$

- On peut *modéliser* de telles données comme :

$$x_i = \mu + \epsilon_i; i = 1, \dots, n;$$

$$y_j = \mu + \Delta + \tau_j; j = 1, \dots, m,$$

où Δ signifie l'effet du traitement (par rapport au groupe 'contrôle')

- $H : \Delta = 0$ vs. $A : \Delta \neq 0$ ou $A : \Delta > 0$ ou $A : \Delta < 0$

Variances égales, cont.

■ $T = \text{diff. observée.} / \text{ES}(\text{diff. observée.}) = \frac{\Delta}{\sqrt{\text{Var}(\hat{\Delta})}} ;$

$$\hat{\Delta} = \bar{y} - \bar{x} ; \text{Var}(\hat{\Delta}) = \frac{\sigma^2}{n} + \frac{\sigma^2}{m} = \frac{n+m}{nm} \sigma^2$$

- On suppose que :

- les variances des 2 échantillons sont *égales* :

$$\text{Var}(\epsilon) = \text{Var}(\tau)$$

- les observations sont *indépendantes*

- *les 2 échantillons sont indépendants*

- On peut estimer les variances *séparément* :

$$s_x^2 = ((x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})^2) / (n - 1)$$

$$s_y^2 = ((y_1 - \bar{y})^2 + \dots + (y_m - \bar{y})^2) / (m - 1)$$

- Quand les variances sont *égales*, on peut combiner les deux estimateurs : $s_p^2 = ((n-1)s_x^2 + (m-1)s_y^2) / (n+m-2)$

$$\Rightarrow t_{obs} = \frac{\bar{y} - \bar{x}}{\sqrt{s_p^2(n+m)/(nm)}} \sim t_{n+m-2} \text{ sous } H$$

Test de comparaison de 2 moyennes : variances inégales

- Si $\sigma_x^2 \neq \sigma_y^2$, on peut utiliser

$$T_{Welch} = \frac{\bar{Y} - \bar{X}}{\sqrt{S_x^2/n + S_y^2/m}}$$

- La distribution de cette statistique T_{Welch} *n'est qu'approximativement* t , avec un nombre de degrés de liberté calculé à la base de s_x , s_y , n et m
- Welch test
- Dans la pratique, si les variances sont assez différentes (rapport plus de 3), on utilise cette statistique (au lieu de celle avec la variance s_p^2)

Exemple

Exemple 9.2

Dépenses d'énergie dans les groupes de femmes minces et obèses :

mince	7.53	7.48	8.08	8.09	10.15	8.40	10.88	6.13	7.90	7.05	7.48	7.58	8.11
obese	9.21	11.51	12.79	11.85	9.97	8.79	9.69	9.68	9.19				

- Tester l'hypothèse que les moyennes des deux populations sont égales ...

Test

1

2

3

4

5

Expériences appariées

- Pour une expérience effectuée en *blocs de deux unités*, la *puissance* du *t*-test pourrait être augmentée
- Cette idée permet *d'éliminer les influences d'autres variables* (p. ex. l'âge, le sexe, etc.), en leur donnant des 'traitements' différents
- Ainsi, on a une comparaison des deux conditions *plus précise*

t-test pour une expérience appariée

- Les données sont de forme :

	1	2	...	n	
contrôle	x_1	x_2	$\cdots$	x_n	espérance μ
traitement	y_1	y_2	$\cdots$	y_n	espérance $\mu + \Delta$

- *Chaque bloc* nous permet d'évaluer l'effet du traitement
- En effet, on considère *les différences*

$$d_1 = y_1 - x_1, \dots, d_n = y_n - x_n$$

comme un échantillon de mesures provenant d'une distribution d'espérance Δ

- $H: \Delta = 0$ vs. $A: \Delta \neq 0$ ou $A: \Delta > 0$ ou $A: \Delta < 0$
- $T = t\text{-apparié} = \frac{\bar{d}}{s_d/\sqrt{n}}$, où
$$s_d^2 = ((d_1 - \bar{d})^2 + \dots + (d_n - \bar{d})^2)/(n-1)$$
- Sous H , $t\text{-apparié} \sim t_{n-1}$

Exemple 9.1, cont.

Prise quotidienne d'énergie des 11 femmes pré- et post-menstruel :

pré	5260	5470	5640	6180	6390	6515	6805	7515	7515	8230	8770
post	3910	4220	3885	5160	5645	4680	5265	5975	6790	6900	7335

- Tester l'hypothèse qu'il n'y a pas de différence de prise quotidienne pré et post ...

Test

1

2

3

4

5