UNIVERSITÉ
DE GENÈVE

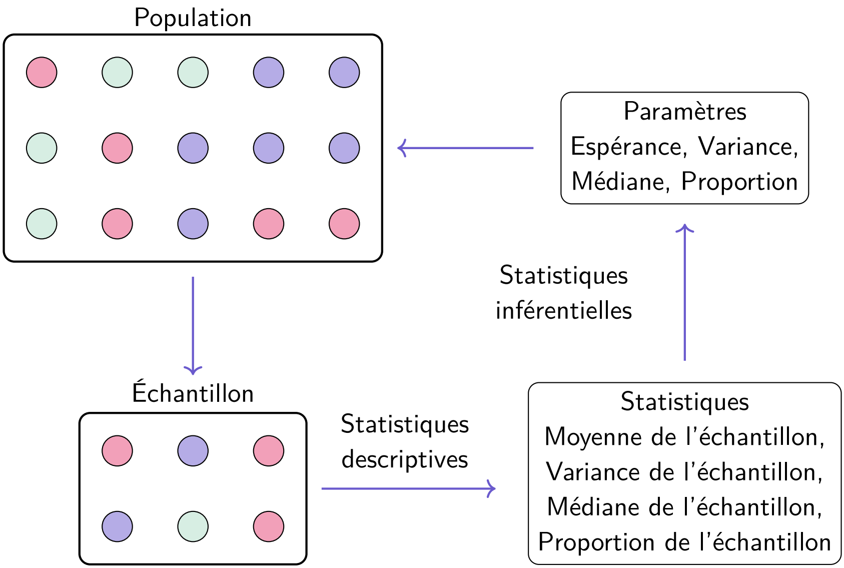
Introduction à la Statistique

Inférence statistique: test-Z, test-t et test-Z à deux échantillons

Stéphane Guerrier, Mucyo Karemera, Samuel Orso & Lionel Voirol

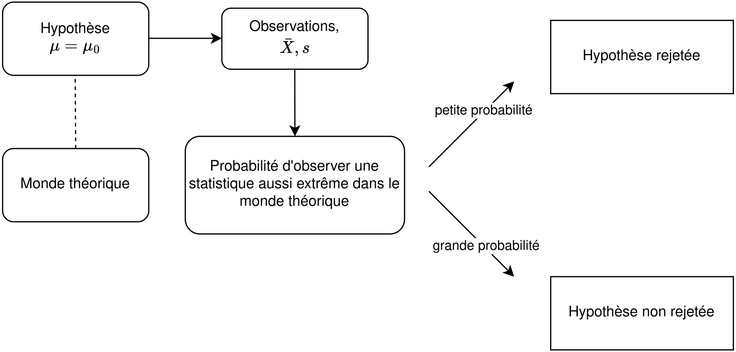
Data Analytics Lab

Inférence statistique



Test d'hypothèse

Est-il possible de tester que l'espérance μ d'une variable aléatoire X soit égale à une certaine valeur μ_0 ?



Exemple : test-Z

Une entreprise de production de batteries affirme que la durée de vie moyenne de ses batteries de type A est de 100 heures.

Un ingénieur souhaite vérifier cette affirmation. Il prélève un échantillon aléatoire de 100 batteries de type A et observe une durée de vie moyenne de 99.1 heures. On considère la durée de vie de chaque batterie comme indépendante.

Il estime ensuite l'écart-type de la population avec l'écart-type de l'échantillon s qui est de 5

Il souhaite déterminer, au seuil de signification de 5%, si les données fournissent des preuves que la durée de vie moyenne des batteries de type A diffère de 100, ou si l'écart observé peut simplement être dû au hasard.

Exemple : test-Z

On considère X : la durée de vie d'une batterie de type A

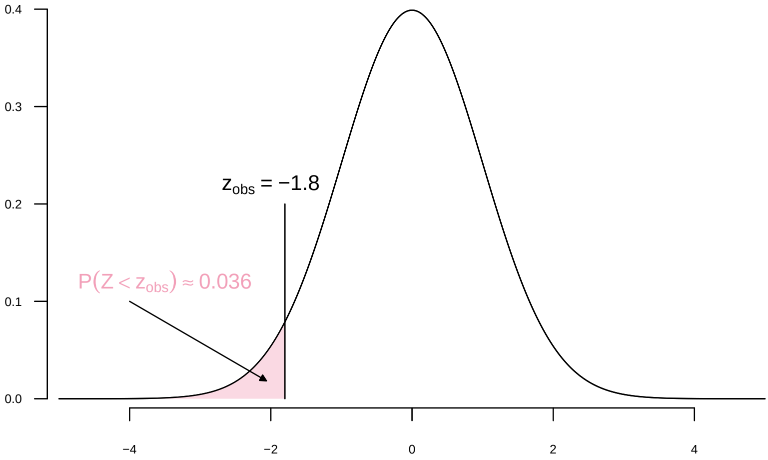
On indique d'abord les hypothèses :

$$H_0 : \mu = \mu_0 = 100$$
$$H_\alpha : \mu \neq 100$$

On calcule ensuite la valeur z_{obs}

$$z_{\text{obs}} = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} = \frac{99.1 - 100}{5/\sqrt{100}} = -1.8$$

Exemple : test-Z



Exemple : test-Z

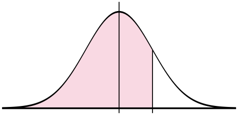
On calcule ensuite la *p*-valeur associée au test étant donné l'hypothèse alternative

H_α	*p*-valeur
$\mu > \mu_0$	$p = \mathbb{P}(Z \geq z_{\text{obs}}) = 1 - \mathbb{P}(Z \leq z_{\text{obs}})$
$\mu < \mu_0$	$p = \mathbb{P}(Z \leq z_{\text{obs}})$
$\mu \neq \mu_0$	$p = \mathbb{P}(\|Z\| \geq \|z_{\text{obs}}\|) = 2\mathbb{P}(Z \geq \|z_{\text{obs}}\|)$

La *p*-valeur est donc de

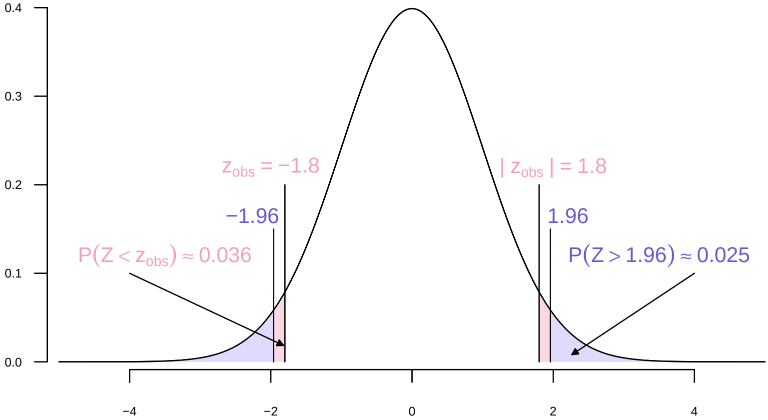
$$\begin{aligned} p &= \mathbb{P}(|Z| \geq |z_{\text{obs}}|) = 2 \times \mathbb{P}(Z \geq |z_{\text{obs}}|) \\ &= 2 \times \mathbb{P}(Z \geq |-1.8|) = 2 \times \mathbb{P}(Z \geq 1.8) \\ &= 2 \times \{1 - \mathbb{P}(Z \leq 1.8)\} \approx 0.0718 \end{aligned}$$

Table de la distribution cumulative de Z



Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817

Exemple : test-Z



Conclusion : À un niveau de signification de 5%, on ne peut pas rejeter l'hypothèse nulle $H_0 : \mu = 100$ au profit de $H_\alpha : \mu \neq 100$ puisque $0.0718 > 0.05$

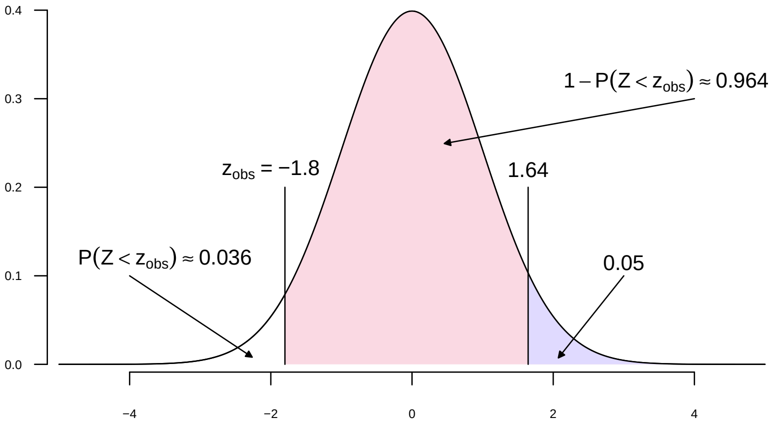
Exemple : test-Z

Un second ingénieur pense que la moyenne de la population est supérieure à 100 heures et souhaite tester les hypothèses suivantes :

$$H_0 : \mu = \mu_0 = 100$$
$$H_\alpha : \mu > 100$$

La p-valeur est donc de $1 - \mathbb{P}(Z \leq z_{\text{obs}}) \approx 1 - 0.0359 \approx 0.964$

Exemple : test-Z



Conclusion : À un niveau de signification de 5%, on ne peut pas rejeter l'hypothèse nulle $H_0 : \mu = 100$ au profit de $H_\alpha : \mu > 100$ puisque $0.964 > 0.05$

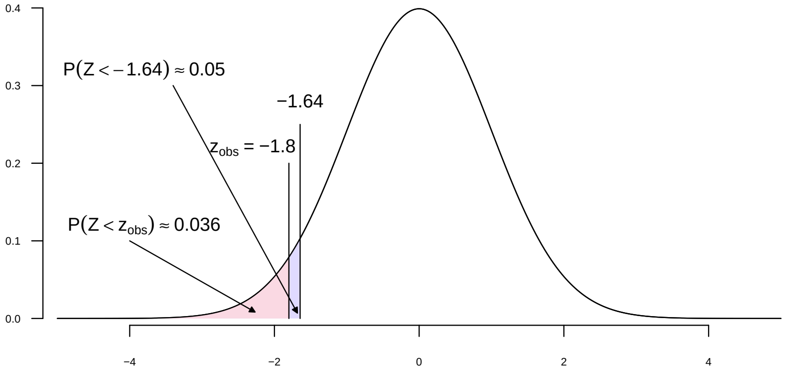
Exemple : test-Z

Un troisième ingénieur pense que la moyenne de la population est inférieure à 100 heures et souhaite tester les hypothèses suivantes :

$$H_0 : \mu = \mu_0 = 100$$
$$H_\alpha : \mu < 100$$

La p-valeur est donc de $\mathbb{P}(Z \leq z_{\text{obs}}) \approx 0.0359$

Test-Z



Conclusion : À un niveau de signification de 5%, on peut rejeter l'hypothèse nulle $H_0 : \mu = 100$ au profit de $H_\alpha : \mu < 100$ puisque $0.0359 < 0.05$

Exemple : Test-t

La même entreprise de production de batteries affirme que la durée de vie moyenne de ses batteries de type B est de 50 heures.

Un ingénieur souhaite vérifier cette affirmation. Il prélève un échantillon aléatoire de 10 batteries de type B et observe une durée de vie moyenne de 47.1 heures. On considère la durée de vie de chaque batterie de type B comme indépendante et distribuée normalement

Il estime ensuite l'écart-type de la population avec l'écart-type de l'échantillon s qui est de 4.5

Il souhaite déterminer, au seuil de signification de 5%, si les données fournissent des preuves que la durée de vie moyenne des batteries de type B diffère de 50, ou si l'écart observé peut simplement être dû au hasard.

Exemple : Test-t

On considère X : la durée de vie d'une batterie de type B

On indique d'abord les hypothèses :

$$H_0 : \mu = \mu_0 = 50$$
$$H_\alpha : \mu \neq 50$$

On calcule ensuite la valeur t_{obs}

$$t_{\text{obs}} = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} = \frac{47.1 - 50}{4.5/\sqrt{10}} \approx -2.038$$

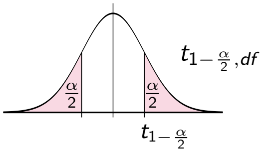
Exemple : Test-t

On cherche ensuite la valeur critique en fonction de α, H_α et n afin de comparer la valeur t_{obs} à la valeur critique

Hypothèse alternative	**Hypothèses**	**Règle de Décision**
Bilatérale	$H_0 : \mu = \mu_0$ $H_\alpha : \mu \neq \mu_0$	Rejetez H_0 si $\lvert t_{\text{obs}} \rvert > t_{1-\frac{\alpha}{2},df}$
Unilatéral (droite)	$H_0 : \mu = \mu_0$ $H_\alpha : \mu > \mu_0$	Rejetez H_0 si $t_{\text{obs}} > t_{1-\alpha,df}$
Unilatéral (gauche)	$H_0 : \mu = \mu_0$ $H_\alpha : \mu < \mu_0$	Rejetez H_0 si $t_{\text{obs}} < -t_{1-\alpha,df}$

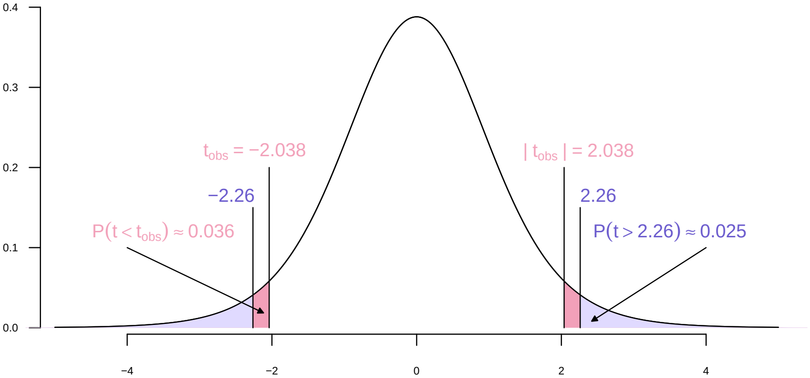
ou $df = n - 1$

Table de la loi de Student



df \ α	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.05	0.02	0.01
1	0.1584	0.3249	0.5095	0.7265	1.0000	1.3764	1.9626	3.0777	6.3138	12.7062	31.8205	63.6567
2	0.1421	0.2887	0.4447	0.6172	0.8165	1.0607	1.3862	1.8856	2.9200	4.3027	6.9646	9.9248
3	0.1366	0.2767	0.4242	0.5844	0.7649	0.9785	1.2498	1.6377	2.3534	3.1824	4.5407	5.8409
4	0.1338	0.2707	0.4142	0.5686	0.7407	0.9410	1.1896	1.5332	2.1318	2.7764	3.7469	4.6041
5	0.1322	0.2672	0.4082	0.5594	0.7267	0.9195	1.1558	1.4759	2.0150	2.5706	3.3649	4.0321
6	0.1311	0.2648	0.4043	0.5534	0.7176	0.9057	1.1342	1.4398	1.9432	2.4469	3.1427	3.7074
7	0.1303	0.2632	0.4015	0.5491	0.7111	0.8960	1.1192	1.4149	1.8946	2.3646	2.9980	3.4995
8	0.1297	0.2619	0.3995	0.5459	0.7064	0.8889	1.1081	1.3968	1.8595	2.3060	2.8965	3.3554
9	0.1293	0.2610	0.3979	0.5435	0.7027	0.8834	1.0997	1.3830	1.8331	2.2622	2.8214	3.2498
10	0.1289	0.2602	0.3966	0.5415	0.6998	0.8791	1.0931	1.3722	1.8125	2.2281	2.7638	3.1693
11	0.1286	0.2596	0.3956	0.5399	0.6974	0.8755	1.0877	1.3634	1.7959	2.2010	2.7181	3.1058
12	0.1283	0.2590	0.3947	0.5386	0.6955	0.8726	1.0832	1.3562	1.7823	2.1788	2.6810	3.0545
13	0.1281	0.2586	0.3940	0.5375	0.6938	0.8702	1.0795	1.3502	1.7709	2.1604	2.6503	3.0123
14	0.1280	0.2582	0.3933	0.5366	0.6924	0.8681	1.0763	1.3450	1.7613	2.1448	2.6245	2.9768
15	0.1278	0.2579	0.3928	0.5357	0.6912	0.8662	1.0735	1.3406	1.7531	2.1314	2.6025	2.9467
16	0.1277	0.2576	0.3923	0.5350	0.6901	0.8647	1.0711	1.3368	1.7459	2.1199	2.5835	2.9208
17	0.1276	0.2573	0.3919	0.5344	0.6892	0.8633	1.0690	1.3334	1.7396	2.1098	2.5669	2.8982
18	0.1274	0.2571	0.3915	0.5338	0.6884	0.8620	1.0672	1.3304	1.7341	2.1009	2.5524	2.8784
19	0.1274	0.2569	0.3912	0.5333	0.6876	0.8610	1.0655	1.3277	1.7291	2.0930	2.5395	2.8609
20	0.1273	0.2567	0.3909	0.5329	0.6870	0.8600	1.0640	1.3253	1.7247	2.0860	2.5280	2.8453

Exemple : Test-t



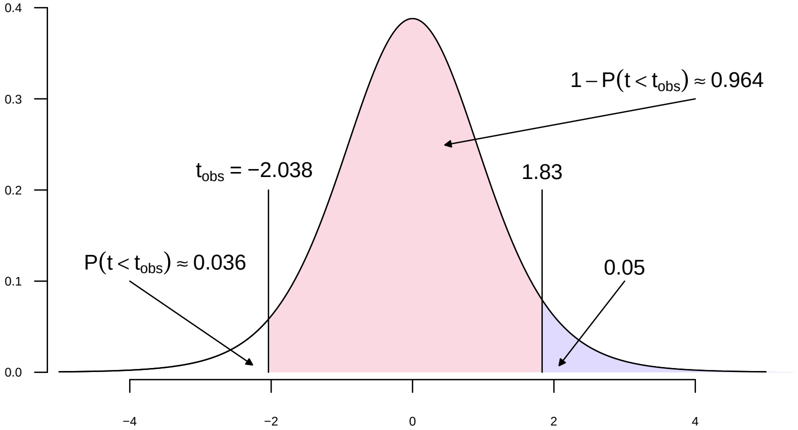
Conclusion : À un niveau de signification de 5%, on ne peut pas rejeter l'hypothèse nulle $H_0 : \mu = 50$ au profit de $H_\alpha : \mu \neq 50$ puisque $2.038 < 2.26$

Exemple : Test-t

Un second ingénieur pense que la moyenne de la population est supérieure à 50 heures et souhaite tester les hypothèses suivantes :

$$H_0 : \mu = \mu_0 = 50$$
$$H_\alpha : \mu > 50$$

Exemple : Test-t



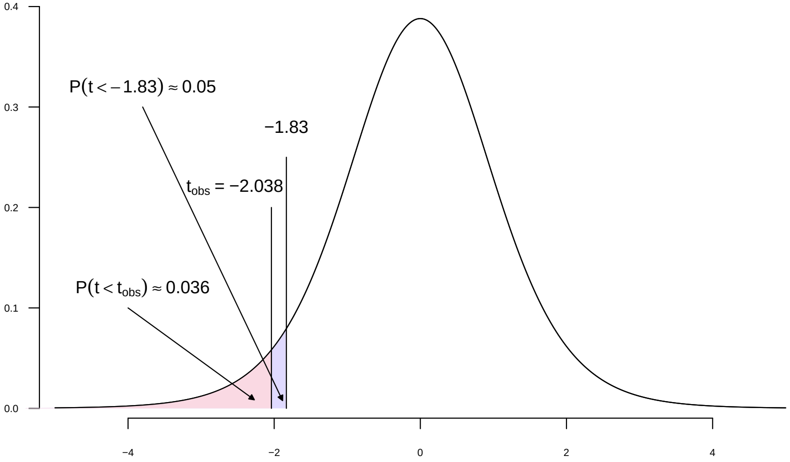
Conclusion : À un niveau de signification de 5%, on ne peut pas rejeter l'hypothèse nulle $H_0 : \mu = 50$ au profit de $H_\alpha : \mu > 50$ puisque $-2.038 < 1.83$

Exemple : Test-t

Un troisième ingénieur pense que la moyenne de la population est inférieure à 50 heures et souhaite tester les hypothèses suivantes :

$$H_0 : \mu = \mu_0 = 50$$
$$H_\alpha : \mu < 50$$

Exemple : Test-t



Conclusion : À un niveau de signification de 5%, on peut rejeter l'hypothèse nulle H_0 : $\mu = 50$ au profit de H_α : $\mu < 50$ puisque $-2.038 < -1.83$

Tests à deux échantillons

Il est courant de vouloir déterminer des différences entre deux groupes. Par exemple,

- En médecine : comparer un traitement à un placebo
- En économie : comparer deux politiques publiques
- Dans l'industrie : comparer les performances entre deux machines

Objectif statistique : déterminer si les différences observées sont dues au hasard ou à un effet réel.

On considère deux échantillons indépendants, chacun provenant d'une des deux populations. Ces échantillons ne sont pas nécessairement de même taille.

On fait un test pour déterminer si les moyennes des populations sont les mêmes.

Exemple

L'entreprise de production de batteries possède deux chaînes de production pour les batteries de types A. On voudrait tester si la durée de vie des batteries diffère en fonction de la chaîne de production

Un échantillon aléatoire de 80 batteries de la chaîne de production 1 et un échantillon aléatoire de 70 batteries de la chaîne de production 2 sont prélevés
On suppose que les observations sont IID et que les groupes sont indépendants

Chaîne de production 1 : $\bar{X}_1 = 99.3$ heures, $s_1 = 4.7$, $n_1 = 80$
Chaîne de production 2 : $\bar{X}_2 = 100.8$ heures, $s_2 = 5.4$, $n_2 = 70$

On souhaite déterminer, au niveau de signification de 5%, si les données fournissent des preuves que la durée de vie moyenne des batteries issue de la chaîne de production 1 (μ_1) diffère de celle issue de la chaîne de production 2 (μ_2), ou si l'écart observé peut simplement être dû au hasard

Test-Z à deux échantillons

Populations IID, groupes indépendants, σ inconnue, grands échantillons ($n \geq 30$)

Statistique de test :

$$z_{\text{obs}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Distribution de z_{obs} sous H_0

$$z_{\text{obs}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim \mathcal{N}(0, 1)$$

Règle de décision : Basée sur la p-valeur et le niveau de signification comme pour les tests à un échantillon

Retour sur le dernier exemple

Hypothèses : $H_0 : \mu_1 = \mu_2 \quad H_1 : \mu_1 \neq \mu_2$

Niveau de signification : $\alpha = 5\%$

Statistique de test :

$$z_{\text{obs}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{99.3 - 100.8}{\sqrt{\frac{4.7^2}{80} + \frac{5.4^2}{70}}} \approx -1.8$$

Calcul de la p-valeur :

$$p\text{-valeur} = 2 \times P(Z \geq |-1.8|) = 2 \times \{1 - P(Z \leq 1.8)\} = ... \approx 7.18\%$$

Conclusion : Au niveau de signification de 5%, on ne peut rejeter $H_0 : \mu_1 = \mu_2$, puisque : p-valeur $= 7.18\% > \alpha = 5\%$

Exemple : test d'un médicament

On teste un nouveau médicament pour réduire la pression artérielle

- Groupe 1 : 50 patients reçoivent le médicament
- Groupe 2 : 50 patients reçoivent un placebo
- Moyenne groupe 1 : $\bar{X}_1 = 125$ mmHg, écart-type $s_1 = 10$
- Moyenne groupe 2 : $\bar{X}_2 = 130$ mmHg, écart-type $s_2 = 9.6$

On souhaite tester les hypothèses suivantes :

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 < \mu_2$$

On souhaite déterminer, au niveau de signification de 5%, si les données fournissent des preuves que le médicament diminue la pression artérielle, ou si l'écart observé peut simplement être dû au hasard

Calcul du test-Z unilatéral

Hypothèses : $H_0 : \mu_1 = \mu_2 \quad H_1 : \mu_1 < \mu_2$

Niveau de signification : $\alpha = 5\%$

Données : $\bar{X}_1 = 125,\ \bar{X}_2 = 130,\ s_1 = 10, s_2 = 9.6,\ n_1 = n_2 = 50$

Statistique de test :

$$z_{\text{obs}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{125 - 130}{\sqrt{\frac{10^2}{50} + \frac{9.6^2}{50}}} \approx -2.55$$

p-valeur (test unilatéral à gauche) :

$$P(Z \leq -2.55) \approx 0.0054$$

Conclusion : Au niveau de signification 5%, on rejette $H_0 : \mu_1 = \mu_2$ en faveur de $H_1 : \mu_1 < \mu_2$, puisque : p-valeur $\approx 0.0054 < 0.05$

Randomisation

Reprenons l'exemple du médicament et supposons que

- le choix de recevoir le traitement ait été fait directement par la personne concernée
- ou que le traitement ait été prescrit en priorité aux situations les plus graves par le personnel soignant

Problème : les groupes pourraient différer systématiquement avant même le traitement

Le médicament semble efficace, mais cet effet pourrait venir d'un biais de sélection

Solution : l'assignation aléatoire (randomisation) garantit que les différences observées sont dues, en moyenne, au traitement

Que faire si les échantillons sont petits ?

Lorsque les échantillons sont petits, et que la normalité est douteuse, d'autres tests peuvent être utilisés

Par exemple : le test-t à deux échantillons, le test de Welch (en cas de variances inégales), ou le test de Wilcoxon (non paramétrique)

Ces tests ne sont pas abordés ici, mais l'idée générale reste la même : utiliser une statistique de test pour décider si les données sont compatibles avec l'hypothèse nulle H_0

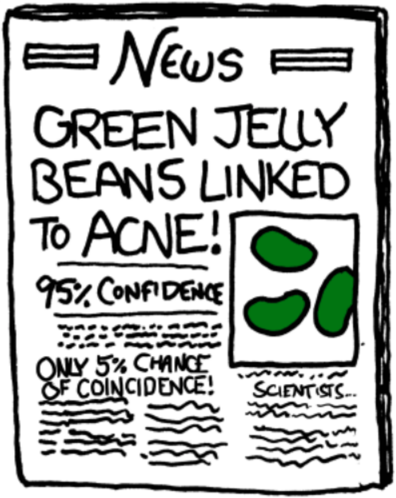
D’autres tests à deux échantillons ? ou plus ?

On peut aussi effectuer des tests à deux échantillons pour

- des proportions
- des variances

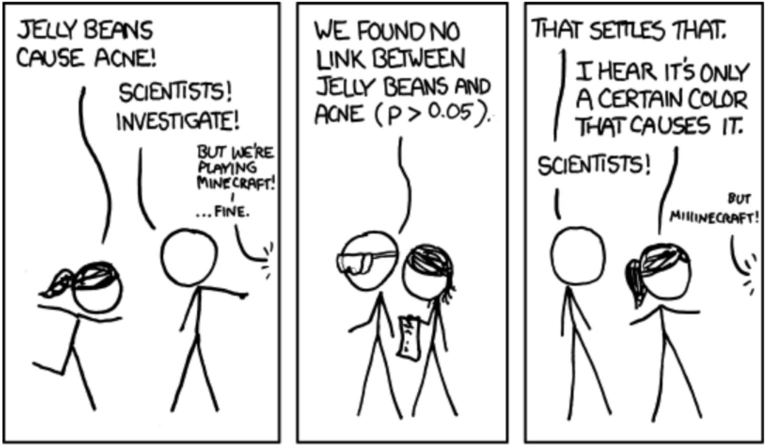
Il aussi possible d’effectuer des tests pour comparer les moyennes de plus de deux échantillons/groupes, mais... attention ! !

Exemple : les jelly beans provoquent-ils l'acné ?



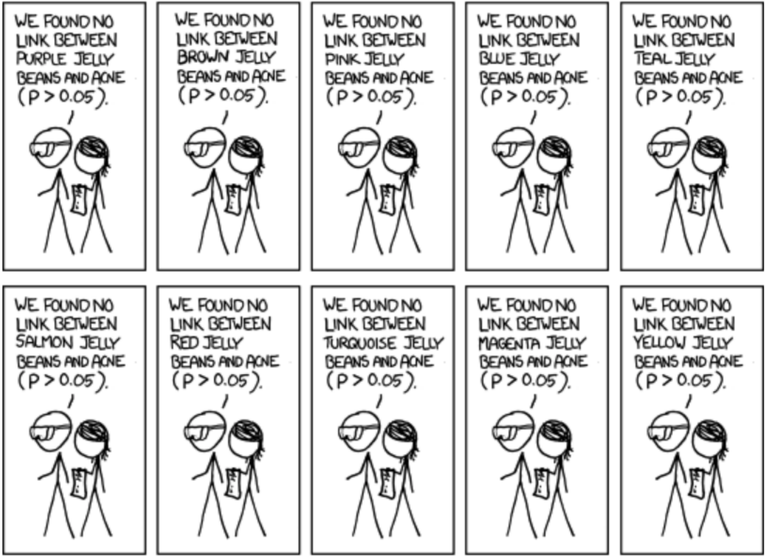
Source : xkcd

Exemple : les jelly beans provoquent-ils l'acné ?



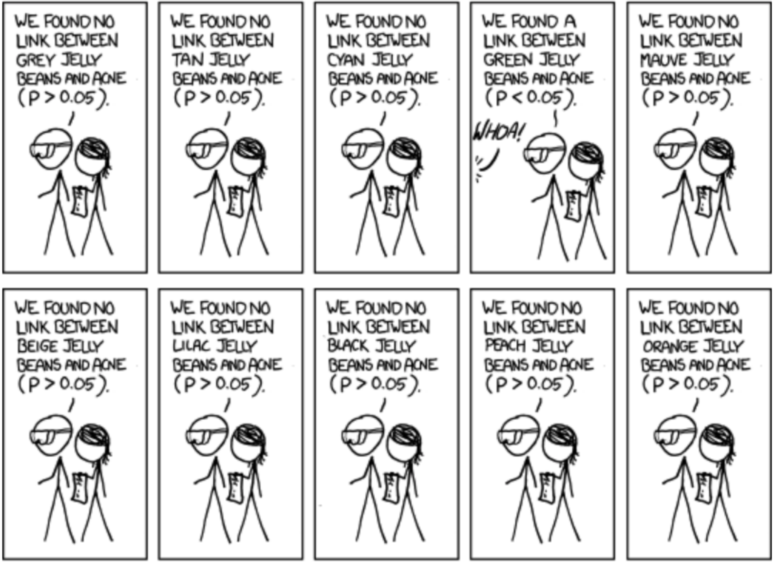
Source : xkcd

Exemple : peut-être une couleur en particulier ?



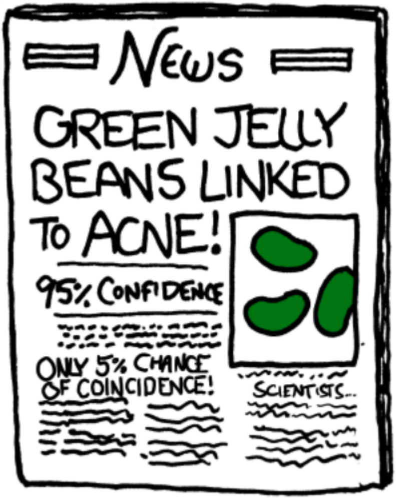
Source : xkcd

Exemple : peut-être une couleur en particulier ?



Source : xkcd

Exemple : finalement...



Source : xkcd

Multiplier les tests peut être dangereux !

La p-valeur est une quantité aléatoire, car elle dépend des données échantillonnées

Il peut donc arriver, par hasard, qu'une p-valeur soit inférieure au niveau de signification choisi, même si H_0 est vraie et ne devrait pas être rejetée. Au niveau de signification de 5%, cela signifie qu'en moyenne, même sans effet réel, la p-valeur sera inférieure à 5% une fois sur vingt

Lorsque plusieurs hypothèses sont testées, la probabilité d'observer un tel événement (rare) augmente. Ainsi le risque de commettre une erreur de type I, i.e., de rejeter à tort H_0, augmente lui aussi

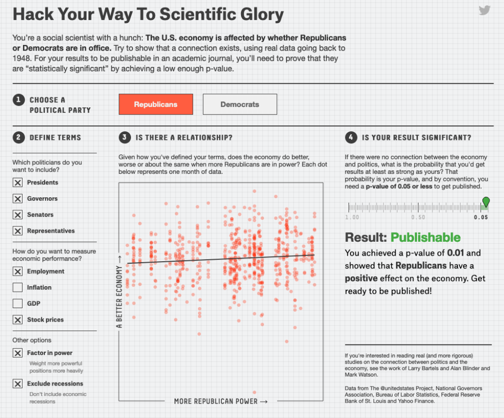
Faire plusieurs tests, sur les mêmes données ou sur des données différentes, est donc risqué (et pourtant très courant), car cela produit presque automatiquement des résultats "significatifs" même en l'absence d'effet réel

Des solutions pratiques existent, comme l'ajustement du niveau de signification en fonction du nombre de tests effectués

p-hacking

Le *p*-hacking désigne le fait de chercher des résultats significatifs en multipliant les tests statistiques sur les mêmes données, puis en ne retenant que ceux associés à des *p*-valeurs suffisamment petites

Exemple : les Republicains sont-ils bons ou mauvais pour l'économie ? Les deux ! !



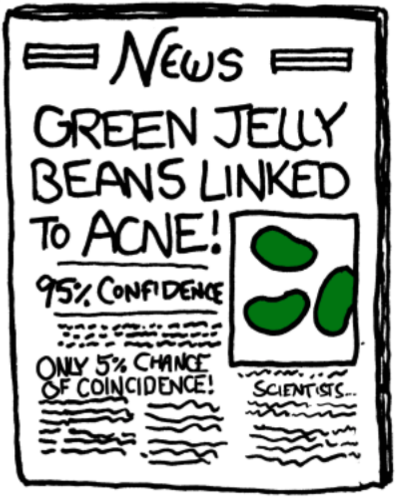
Source : Hack your way to scientific glory

HARKing

Le HARKing est l'acronyme de "Hypothesizing After the Results are Known" désignant les pratiques de recherche questionnables. Cette dérive peut être définie comme le fait de faire passer des hypothèses a posteriori comme des hypothèses a priori

En pratique ?

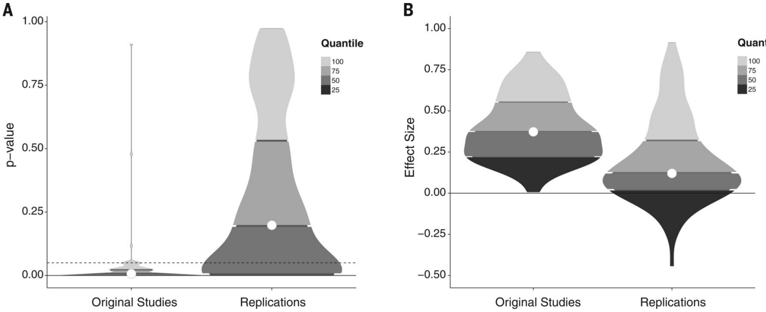
1. Tester si une couleur de jelly bean est liée à l'acné (tests multiples)
2. Prétendre que c'était l'hypothèse originale
3. Publier cette découverte majeure ! ! !



Source : xkcd

Conséquence du *p*-hacking & du HARKing

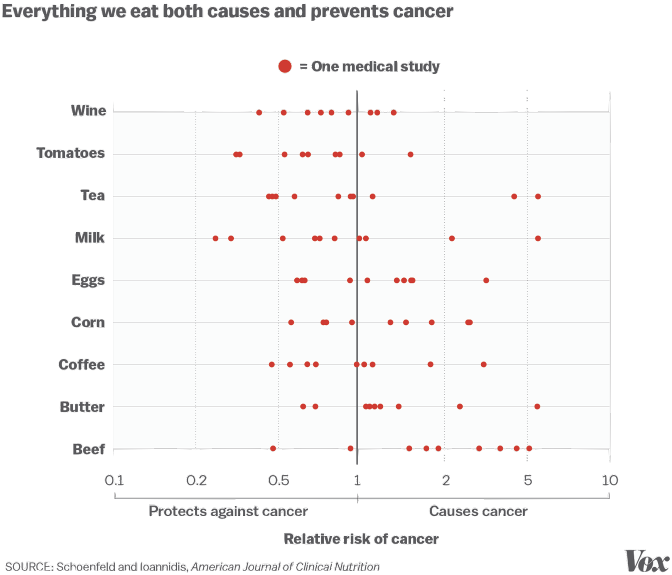
Le *p*-hacking et le HARKing jouent un role prépondérant dans la crise de reproducibilité en science !



Source : Open Science Collaboration. "Estimating the reproducibility of psychological science." Science 349.6251 (2015)

Plus d'informations sur la crise de reproducibilité en science se trouvent ici

Conséquence du p-hacking & du HARKing



Source : "This is why you shouldn't believe that exciting new medical study"

Pour conclure

Toute analyse basée sur des échantillons de données est soumise à de l'aléatoire

Avec les données, il n'y a jamais de certitude à 100 %, mais plutôt une conclusion associée à un risque (estimé) que cette conclusion soit erronée

Il est nécessaire de s'y habituer, et tout résultat scientifique affirmant une validité à 100 % sur la base de données est tout simplement incorrect

En revanche, bien contrôler le risque statistique et formuler correctement les conclusions que l'on peut tirer d'une analyse de données peut réellement permettre d'apporter de nouvelles connaissances, en particulier dans toutes les sciences fondées sur des observations/experimentations

Merci pour votre attention

Bonne chance pour vos examens