Fluctuations d'échantillonnage
Quelques rappels
Intervalle de confiance
Les tests d'hypothèse

Fluctuations d'échantillonnage et tests d'hypothèse

Yohann.Foucher@univ-nantes.fr

Equipe d'Accueil 4275 "Biostatistique, recherche clinique et mesures subjectives en santé", Université de Nantes

Master 2 - Bioinformatique, 26 Octobre 2011

UNIVERSITÉ DE NANTES
CENTRE HOSPITALIER UNIVERSITAIRE DE NANTES
itun institut transplantation urologie néphrologie INSERM - UMR 643

Plan

1. Fluctuations d'échantillonnage

2. Quelques rappels

3. Intervalle de confiance

4. Les tests d'hypothèse

Plan

1. Fluctuations d'échantillonnage

2. Quelques rappels

3. Intervalle de confiance

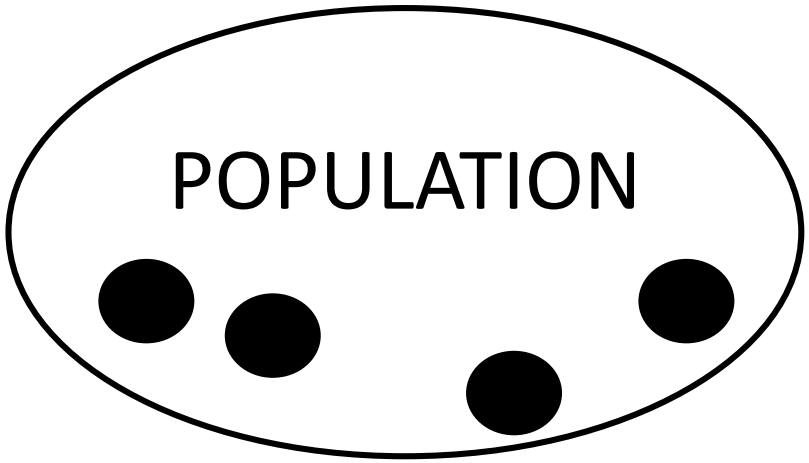
4. Les tests d'hypothèse

Fluctuations d'échantillonnage

- Le réel est monstrueux. Il est énorme, il est hors norme. *
- Etudier totalement une population reviendrait à vouloir percevoir de manière simultanée et continue, toutes les caractéristiques de tous les individus de cette population. †
- Pour approcher et distinguer les choses, pour s'en faire une idée tangible, il faut s'en tenir à l'appréhension d'un nombre limité de caractéristiques.

∗. Edgar Morin
†. Y. Macé. Journal de la Société Française de Statistique, tome 147, 2006.

Fluctuations d'échantillonnage



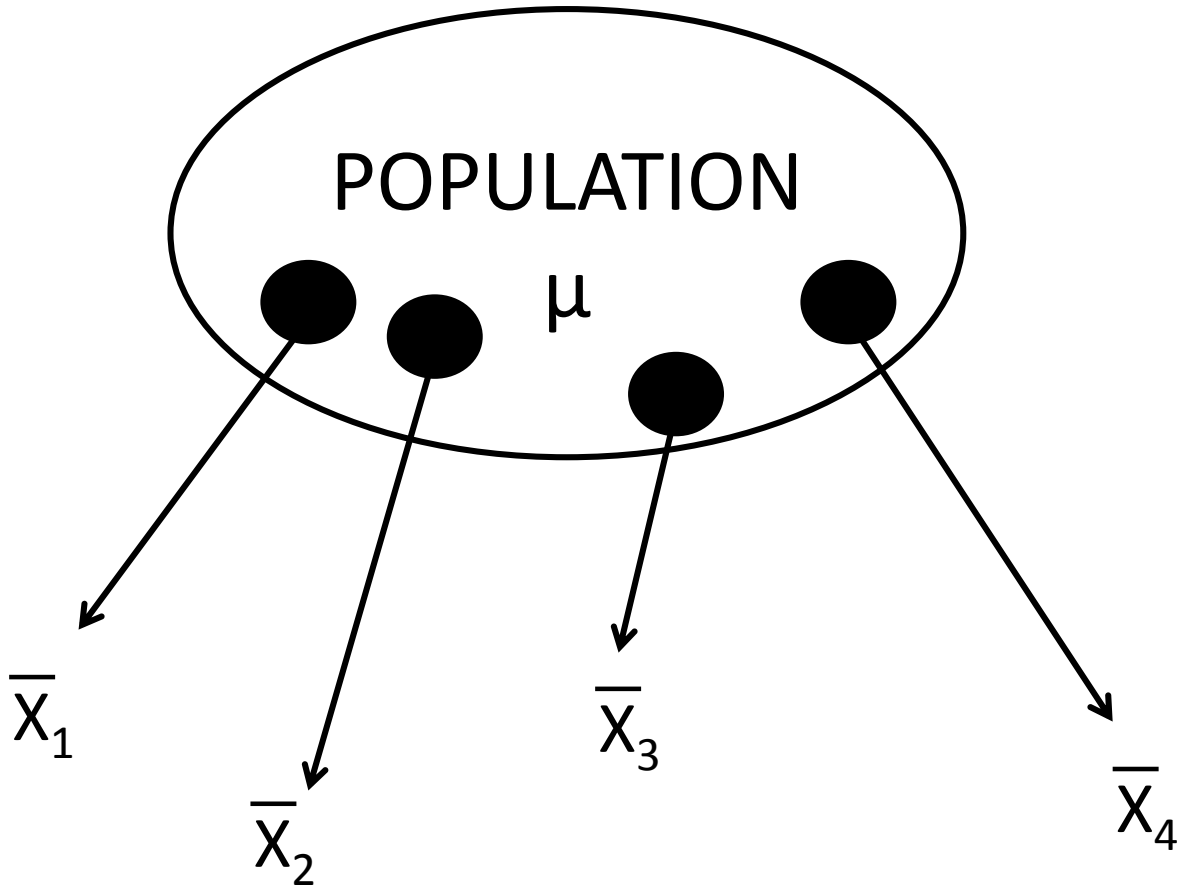
Population non-observable

Echantillons observables

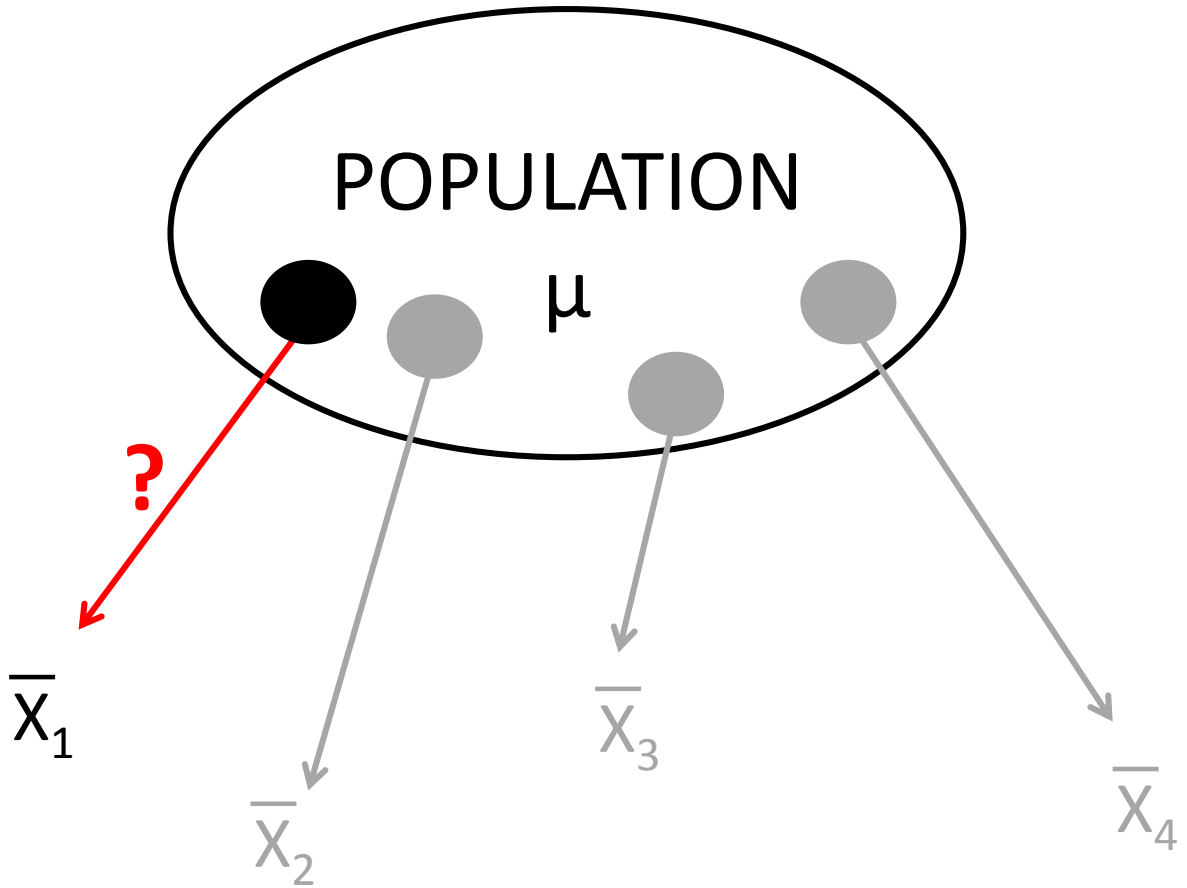
Fluctuations d'échantillonnage

- Soit $\mathcal{P}$ la population exhaustive d'intérêt.
  - Ex : $\mathcal{P}$ = toutes les femmes atteintes d'un cancer du sein.
- Soit μ l'âge moyen de cette population au diagnostic.
  - Ex : $\mu = 55$ ans (supposition).
- Il n'est pas possible de mesurer μ à partir de tous les patients.
- Soit E, un échantillon de N patientes à partir desquelles on observe la moyenne, $\bar{x}$. On espère que $\bar{x}$ est proche de μ.
- Problème : si plusieurs échantillons sont réalisés (E_1, E_2, E_3), on observera autant de moyennes.
  - Ex : $N = 100$ femmes, $\bar{x}_1 = 53.1$, $\bar{x}_2 = 58.0$, $\bar{x}_3 = 56.3$.

Fluctuations d'échantillonnage



Inférence statistique

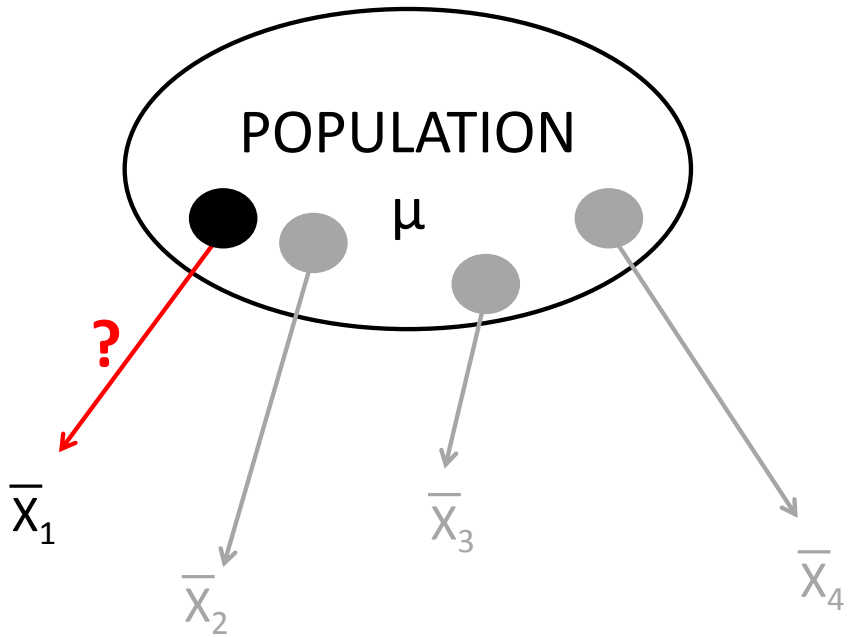


Inférence statistique

- Exemple : On souhaite mesurer l'association entre l'expression d'un gène et le succès thérapeutique d'une chimiothérapie chez les femmes avec un cancer du sein. On mesure chez 100 femmes l'expression par PCR quantitative et on suit les femmes. On sait que l'âge moyen des femmes atteintes d'un cancer du sein est proche de 55 ans. Soit $\bar{x}$ l'âge moyen observé des 100 femmes. Peut-on considérer cet échantillon représentatif ?

a. Si $\bar{x} = 40$ ans : il est très probable que l'échantillon soit composé de femmes plus jeunes qu'en moyenne
   $\rightarrow$ Résultats peu généralisables en pratique
b. Si $\bar{x} = 70$ ans : il est très probable que l'échantillon soit composé de femmes plus âgées qu'en moyenne
   $\rightarrow$ Résultats peu généralisables en pratique
c. Et si $\bar{x} = 58$ ans ?

Inférence statistique



- L'inférence statistique ne conduit pas à une conclusion "stricte".
- Elle attache une probabilité d'erreur à chaque conclusion.

$\Rightarrow$ **AUCUNE CERTITUDE SUR LES CONCLUSIONS**

Plan

1. Fluctuations d'échantillonnage

2. Quelques rappels

3. Intervalle de confiance

4. Les tests d'hypothèse

La notion de variable aléatoire

- Les v.a. ont été définies à l'origine pour représenter un gain.
- Ex : Lancer d'une pièce de monnaie. X peut prendre les valeurs :
  - 1 euro si pile
  - -1 euro si face
- De façon plus générale une variable aléatoire est le résultat inconnu d'une expérience aléatoire.
- Une fois l'experience réalisé, on observe la réalisation de la v.a. X. Cette valeur n'est plus aléatoire mais est certaine. On la note souvent x. x_1 pour la 1ère expérience, x_2 pour la 2nde expérience, etc.

Les principaux estimateurs

- La moyenne de x_1, x_2, ..., x_N :

$$\bar{x} = N^{-1} \sum_{i=1}^{N} x_i$$

- La variance de x_1, x_2, ..., x_N :

$$s^2 = (N-1)^{-1} \sum_{i=1}^{N} (x_i - \bar{x})^2$$

- La proportion de k événements parmi N :

$$p = k/N$$

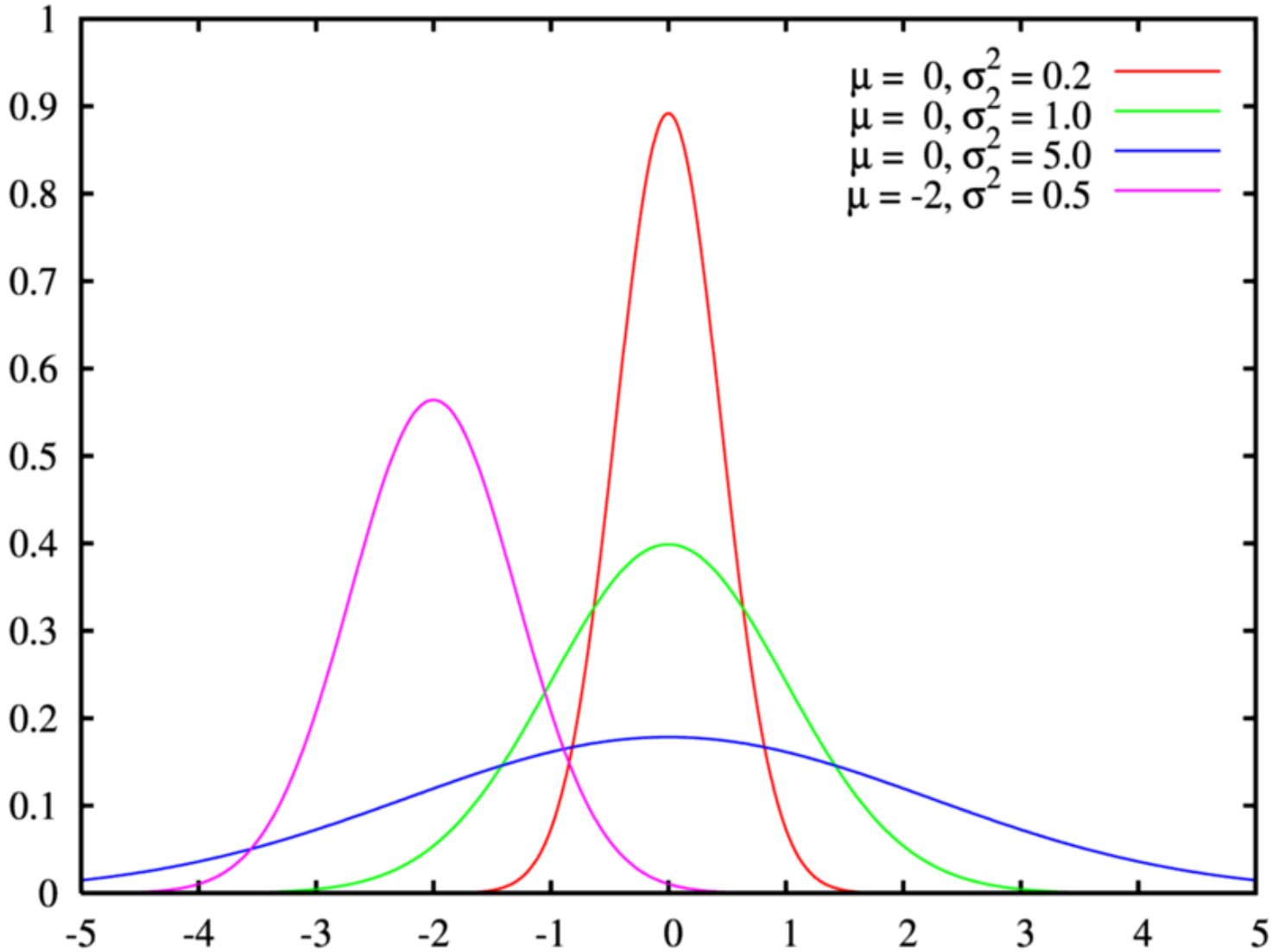
Loi normale

- Loi normale ou loi de Laplace-Gauss
- Caractérisée par sa moyenne (μ) et son écart-type (σ) :

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

- La distribution est symétrique et centrée autour de la moyenne.
- Moyenne = Mode = Médiane.
- L'écart-type représente la dispersion autour de la moyenne.

Loi normale


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-5
-4
-3
-2
-1
0
1
2
3
4
5
μ = 0, σ² = 0.2
μ = 0, σ² = 1.0
μ = 0, σ² = 5.0
μ = -2, σ² = 0.5

Loi normale

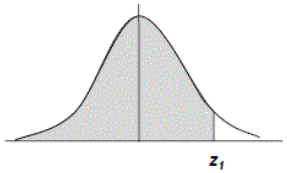
- Si $X \sim \mathcal{N}(\mu, \sigma)$ alors :

$$U = (X - \mu)/\sigma \sim \mathcal{N}(0, 1)$$

.

- $\mathcal{N}(0, 1)$ est la loi normale centrée et réduite.
- En anglais : *standard normal distribution*.
- Intérêt de la transformation : la probabilité $P(T < t)$ est connue.

Loi normale



$$p(z \le z_1) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_1} e^{-\frac{1}{2}z^2} dz$$

z_1	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
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Loi normale

- Symétrie $\rightarrow P(T > t) = 1 - P(T \leq t) = P(T < -t)$.
- Valeurs importantes à retenir :
  - $P(-1,96 < T < 1,96) = 0,95$
  - $P(T < -1,64) = P(T > 1,64) = 0,95$
- Si $X \sim \mathcal{N}(\mu, \sigma)$ alors $\bar{X} \sim \mathcal{N}(\mu, \sigma/\sqrt{n})$

Théorème Centrale Limite (TCL)

- Soit un échantillon de taille N et X_1, X_2, $\cdots$, X_N une suite de variables aléatoires indépendantes et identiquement distribuées.
- Supposons que cette loi possède une espérance μ et un écart-type σ (avec $\sigma \neq 0$).
- Considérons la somme $S_N = X_1 + X_2 + \cdots + X_N = \sum_{i=1}^{N} X_i$. Alors l'espérance de S_N vaut $N\mu$ et son écart-type vaut $\sigma\sqrt{N}$.
- Quand N est assez grand ($N > 30$) , la loi normale $\mathcal{N}(N\mu, N\sigma^2)$ est une bonne approximation de la loi de S_N.
- Cette approximation s'écrit :

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

Théorème Centrale Limite (TCL)

- Conséquence : Une donnée influencée par une multitude de phénomènes aléatoires indépendants qui s'additionnent est approximativement décrite par une loi normale même si les phénomènes qui la composent ne suivent pas des lois normales.

Théorème Centrale Limite (TCL)

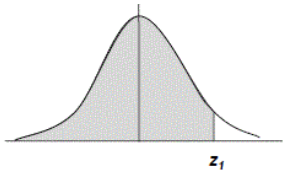
- Exemple

Reprenons l'exemple du cancer du sein. On sait que μ est proche de 55 ans. Supposons l'écart-type est égal à 30 ans.‡ On observe une moyenne $\bar{x}$ de 58 ans, soit 3 ans d'écart avec la valeur théorique. Quelle est la probabilité (A) d'observer un échantillon de 100 femmes avec une moyenne d'âge de plus de 3 ans d'écart avec la moyenne attendue ?

$$\begin{aligned} A &= Pr(\bar{X} > 58) + Pr(\bar{X} < 52) \\ &= Pr(\frac{\bar{X} - 55}{30/\sqrt{100}} > \frac{58 - 55}{30/\sqrt{100}}) + Pr(\frac{\bar{X} - 55}{30/\sqrt{100}} < \frac{52 - 55}{30/\sqrt{100}}) \\ &= Pr(U > 1) + Pr(U < -1) \\ &= 2 \times Pr(U > 1) \end{aligned}$$

‡. Valeur le plus souvent inconnue dans la population et estimée à partir de l'échantillon

Théorème Centrale Limite (TCL)



$$p(z \le z_1) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_1} e^{-\frac{1}{2}z^2} dz$$

z_1	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
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Théorème Centrale Limite (TCL)

$$\begin{aligned} A &= 2 \times Pr(U > 1) \\ &= 2 \times (1 - 0.8413) \\ &= 0.3174 \end{aligned}$$

La probabilité d'observer un échantillon de 100 femmes avec une moyenne d'âge observée écartée de plus de 3 ans avec la moyenne théorique est égale à 0.32. Autrement dit, il y a environ une chance sur trois d'observé un échantillon encore plus éloigné de la population.

- Remarque : On peut penser que l'échantillon observé soit assez vraissemblable et que sa représentativité ne soit pas remise en question.

Approximation d'une loi binomiale

Soit un échantillon de taille N et X_1, X_2, $\cdots$, X_N une suite de variables aléatoires binaires indépendantes et identiquement distribuées selon une loi de Bernoulli de paramètre π.

La somme S_N suit une loi binomiale de paramètre N et π.

Si $N > 30$, $N\pi > 5\%$ et $N(1-p) > 5\%$, alors

$$S_N \sim \mathcal{N}(\mu = N\pi, \sigma = \sqrt{N\pi(1-\pi)})$$

Si P est la proportion estimée à partir de cet échantillon, alors :

$$P = S_N/N \sim \mathcal{N}\left(\mu = \pi, \sigma = \sqrt{\frac{\pi(1-\pi)}{N}}\right)$$

Approximation d'une loi binomiale

- Exemple

On souhaite réalisé un échantillon de taille $N = 100$ individus. Ces individus sont issus d'une population où la prévalence d'un certain phénotype est de 20%. $x_i = 1$ si le sujet i possède ce phénotype et $x_i = 0$ sinon. S_N représente le nombre d'individus avec ce phénotype. Quelle est la probabilité d'observer moins de 10% des individus avec ce phénotype dans l'échantillon ?

$$S_N \sim \mathcal{B}(100, 0.2)$$

Comme $N = 100 > 30$ et $Np = 20 > 5$,

$$S_n \sim \mathcal{N}(20, 4)$$

$$\Downarrow$$

$$P = \frac{S_N}{N} \sim \mathcal{N}(0.20, 0.04)$$

Approximation d'une loi binomiale

$$P \sim \mathcal{N}(0.20, 0.04)$$

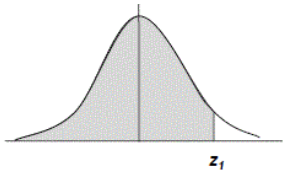
$$\Downarrow$$

$$U = \frac{P - 0.20}{0.04} \sim \mathcal{N}(0, 1)$$

$$\Downarrow$$

$$\begin{aligned} Pr(P < 0.10) &= Pr\left(\frac{P - 0.20}{0.04} < \frac{0.10 - 0.20}{0.04}\right) \\ &= Pr\left(U < \frac{0.10 - 0.20}{0.04}\right) \\ &= Pr(U < -2.50) \end{aligned}$$

Approximation d'une loi binomiale



$$p(z \le z_1) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_1} e^{-\frac{1}{2}z^2} dz$$

z_1	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
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Approximation d'une loi binomiale

$$\begin{aligned} Pr(P < 0.10) &= Pr(U < -2.50) \\ &= 1 - 0.9938 \\ &= 0.0062 \end{aligned}$$

La probabilité d'observer un échantillon de 100 patients parmi lesquels moins de 10% possède le phénotype est égale à 6 pour 1000.

Plan

1. Fluctuations d'échantillonnage

2. Quelques rappels

3. Intervalle de confiance

4. Les tests d'hypothèse

Intervalle de confiance

- Soit un échantillon de taille N et x_1, x_2, $\cdots$, x_N les observations d'une variable aléatoire continue.
- La moyenne de la population (μ) est estimée par $\bar{x}$.
- Si $N > 30$ (TCL), alors on sait aussi que

$$\bar{X} \sim \mathcal{N}(\mu, s/\sqrt{N})$$

- En centrant et réduisant, on obtient :

$$\frac{\bar{X} - \mu}{s/\sqrt{N}} \sim \mathcal{N}(0, 1)$$

- Pour obtenir l'intervalle de confiance à 95% de la moyenne estimée, on cherche l'intervalle à l'intérieur duquel on a 95% de chance de retrouver la moyenne théorique de la population :

$$Pr(-1.96 < \frac{\bar{X} - \mu}{s/\sqrt{N}} < 1.96) = 0.95$$

Intervalle de confiance

$$Pr(-1.96 < \frac{\bar{X} - \mu}{s/\sqrt{N}} < 1.96) = 0.95$$

$$\Leftrightarrow$$

$$Pr(\bar{X} - 1.96\frac{s}{\sqrt{N}} < \mu < \bar{X} + 1.96\frac{s}{\sqrt{N}}) = 0.95$$

$$\Leftrightarrow$$

$$IC_{95\%}(\mu) = [\bar{x} \pm 1.96\frac{s}{\sqrt{N}}]$$

Intervalle de confiance

- Exemple

Reprenons l'exemple du cancer du sein. L'écart-type estimé est de 30 ans. On observe une moyenne $\bar{x}$ de 58 ans. Quelle est l'intervalle de confiance de la moyenne ?

$$\begin{aligned} IC_{95\%}(\mu) &= [58 \pm 1.96 \frac{30}{\sqrt{100}}] \\ &= [52.12; 63.88] \end{aligned}$$

Si on réalise 100 échantillons comme celui-ci ($N = 100$), on attend 95 moyennes estimées comprises entre 52 et 63 ans.

- Remarque : comme la moyenne de la population générale est proche de 55 ans dans l'énoncé, la représentativité de l'échantillon n'est pas remise en question.

Plan

1. Fluctuations d'échantillonnage

2. Quelques rappels

3. Intervalle de confiance

4. Les tests d'hypothèse

Principes

- Principe de la statistique : Est ce que la différence que j'observe est due à la fluctuation d'échantillonnage ?
- J'observe une moyenne $\bar{x}$ (et sa variance) à partir d'un échantillon. Je veux la comparer à une valeur théorique μ (connue dans $\mathcal{P}$). 2 hypothèses :
  - La différence observée est minime. Elle est due au fait que tous les sujets de la population n'ont pas été inclus. Si on avait inclus tout le monde, la moyenne observée $\bar{x}$ serait égale à la moyenne théorique μ.
    - Cette hypothèse est appelée hypothèse nulle : H_0
  - La différence observée est importante. Elle ne peut pas être due au fait que tous les sujets n'ont pas été inclus. Il est évident que si on avait inclus tout le monde, la moyenne observée $\bar{x}$ aurait été différente de la moyenne théorique μ
    - Cette hypothèse est appelée hypothèse alternative : H_1
- Probabilité de me tromper si je rejette H_0 (risque de 1ère espèce).
- Probabilité de me tromper si je rejette H_1 (risque de 2nd espèce).

Principes

- Calculer la probabilité de se tromper en rejetant H_0.
  $\rightarrow$ probabilité critique, notée p_c.
  $\rightarrow$ *p-value* en anglais, notée p.
- C'est la probabilité d'observer l'échantillon en supposant H_0 comme vraie.
  - Si cette probabilité est forte, tendance au non-rejet de H_0.
  - Si cette probabilité est faible, tendance au rejet de H_0.
- On choisi a priori une probabilité maximale d'erreur (α), au delà de laquelle on ne rejettera pas H_0
  - Si $p_c \geq \alpha$: non-rejet de H_0.
  - Si $p_c < \alpha$: rejet de H_0.

Exemple illustratif

- Pour illustrer cette logique, choisissons le cas où l'on souhaite savoir si l'échantillon observé est issu d'une population où la moyenne théorique est égale à μ.
- Exemple : un chercheur souhaite savoir si son échantillon de 100 femmes atteintes d'un cancer du sein est représentatif des femmes atteintes de cette pathologie concernant l'âge au diagnostic ($\mu = 55$ ans). La moyenne estimée à partir de l'échantillon est de 58 ans et la variance estimée est de 900 ans^2 (soit un écart-type de 30 ans).

Exemple illustratif

- Définir la variable aléatoire X dans la population d'étude.
  - $\rightarrow$ Age au diagnostic de cancer : v.a. continue.
- Définir la population de référence $\mathcal{P}$ et la moyenne théorique correspondante μ.
  - $\rightarrow$ Femmes atteintes d'un cancer du sein et $\mu = 55$ ans.
- Choisir les hypothèses à tester :
  - $\rightarrow$ H_0 : $\bar{X} = \mu$, l'échantillon est issu d'une population où l'âge moyen est μ (55 ans).
  - $\rightarrow$ H_1 : $\bar{X} \neq \mu$, l'échantillon est issu d'une population où l'âge moyen n'est pas égal à μ (55 ans).

Exemple illustratif

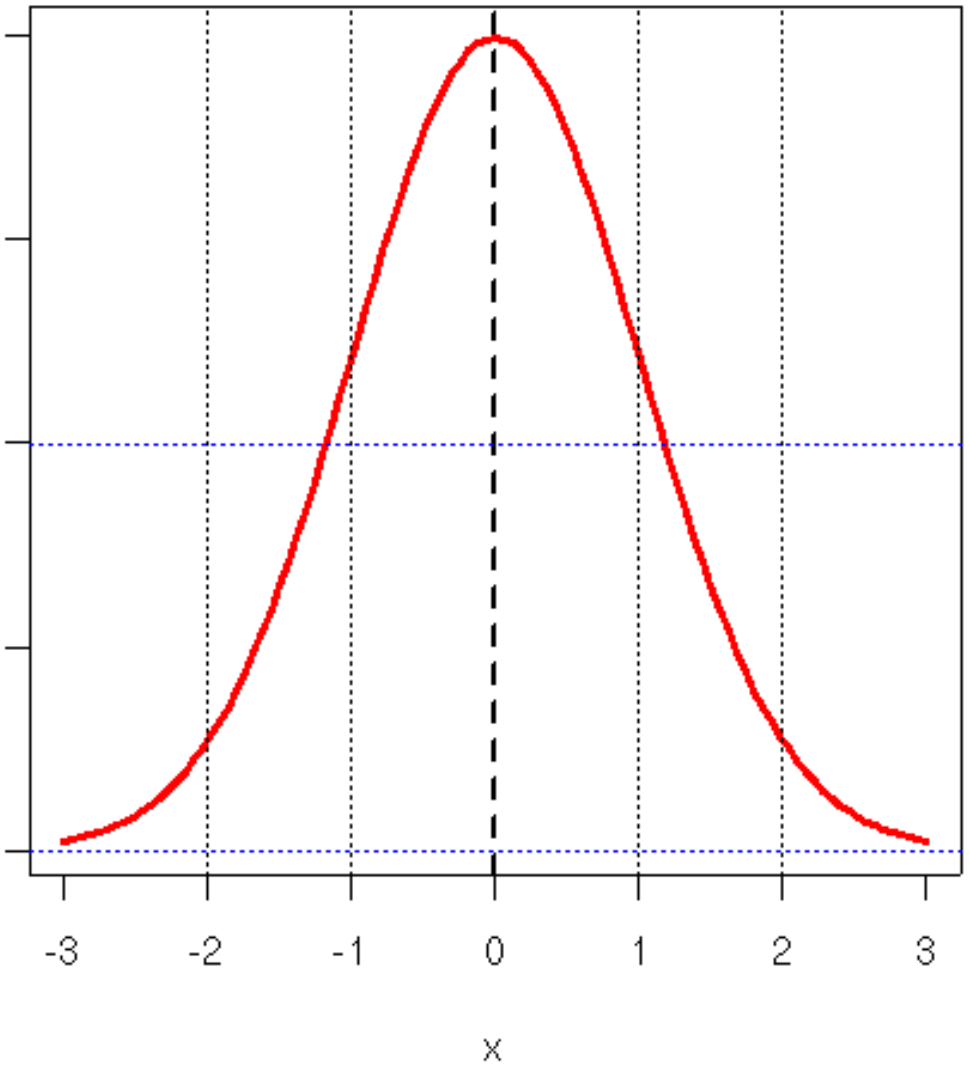
- Définir la statistique de test sous H_0. Comme $N > 30$:

$$\bar{X} \sim \mathcal{N}(\mu, \sigma/\sqrt{N}) \text{ donc } U = \frac{\bar{X} - \mu}{\sigma/\sqrt{N}} \sim \mathcal{N}(0,1)$$

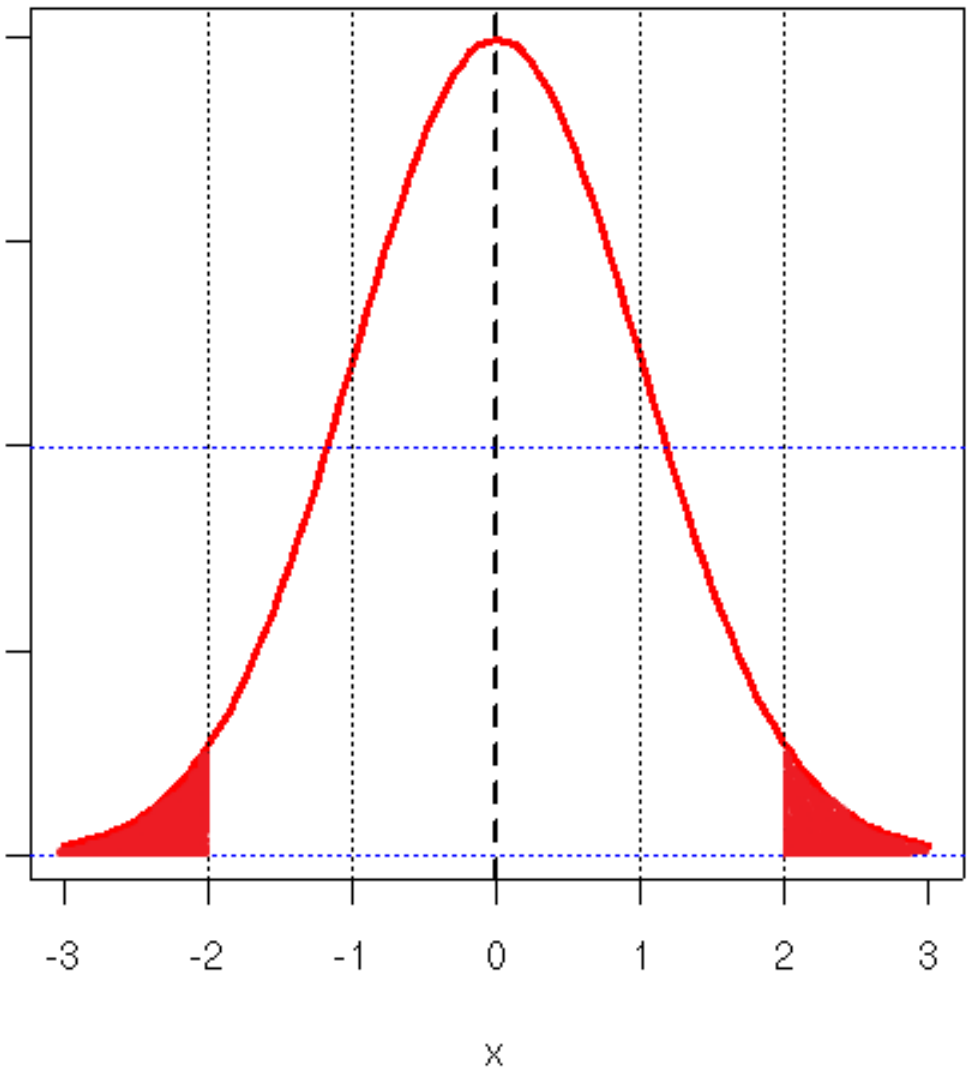
- Définir du risque de 1ère espèce maximum
  - ex : $\alpha = 5\%$
- Définir la région critique (RC). Il s'agit des valeurs de la statistique de test U qui ont moins de α chance d'être observées sous H_0.
  - $\alpha = 5\% \Rightarrow$ Quelle est la valeur de $u_{2.5\%}$ pour que $P(-u_{2.5\%} < U < u_{2.5\%}) = 0.95$?
  - Comme $U \sim \mathcal{N}(0,1) \rightarrow u_{2.5\%} = 1.96 \rightarrow RC : U \notin [-1.96, 1.96]$

  Si la statistique de test est inférieure à -1.96 ou si elle est supérieure à 1.96, l'échantillon observé a moins de 5% de chance de provenir d'une population où H_0 est vraie.

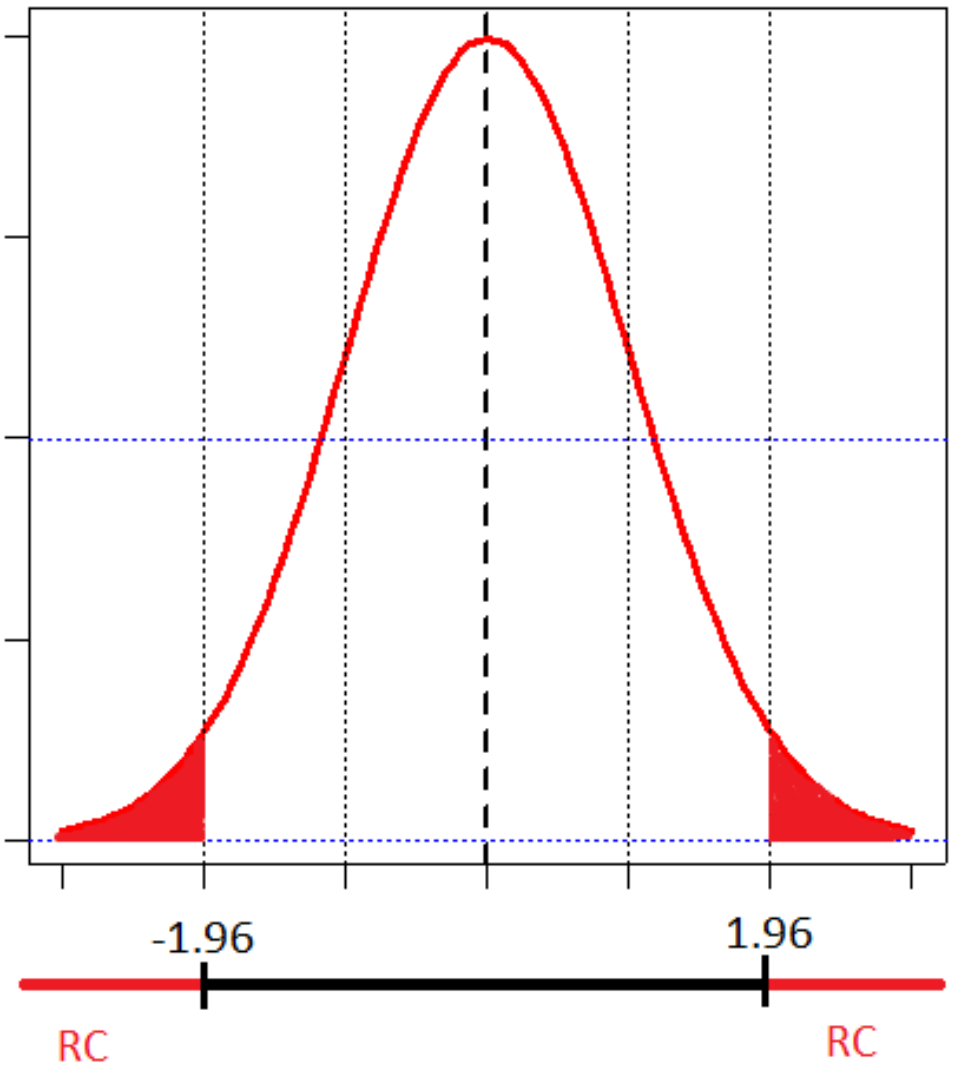
Exemple illustratif



Exemple illustratif

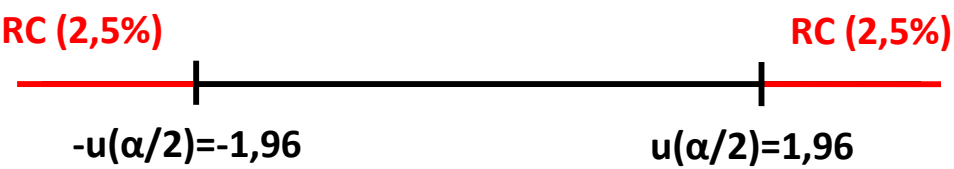


Exemple illustratif



Exemple illustratif

- Application numérique
  - $u = (\bar{x} - \mu)/(s/\sqrt{N})$
  - $N = 100$
  - $\bar{x} = N^{-1} \sum_i x_i = 58$
  - $s^2 = (N-1)^{-1} \sum_i (x_i - \bar{x})^2 = 900$
  - $u = (58 - 55)/(\sqrt{900}/\sqrt{40}) = 1.00$



- Conclusion : La statistique de test n'appartient pas à la région critique. L'hypothèse nulle ne peut donc pas être rejetée. Il n'est pas possible de conclure que l'échantillon ne provienne pas de la population de référence ($p_c > 5\%$).

Remarques Générales

- Le risque d'erreur précédent est de rejeter à tort l'hypothèse alternative
  - Risque de 2nde espèce.
  - Seule interprétation possible : On ne peut pas montrer que l'échantillon n'est pas représentatif.
  - On ne peut pas conclure sur la représentativité.
- Les conclusions ne sont jamais déterministes.
- Incertitude toujours présente et doit ressortir dans la conclusion.
- Si la conclusion avait été le rejet de H_0 car moins de 5% de chance qu'elle soit vraie, la conclusion aurait pu être du type :
  - Il semble que l'échantillon observé ne soit pas représentatif de l'âge moyen de diagnostic du cancer du sein ($p_c < 5\%$).

Stratégie alternative

- Calculer directement la probabilité critique.
- Quelle est la probabilité d'observer cet échantillon ou un échantillon encore moins probable si H_0 est vrai.
- Exemple précédent : $u = 1.00$.
- $p_c = P(U > 1.00) + P(U < -1.00) = 2 * P(U > 1.00) = 2 * (1 - P(U < 1.00))$.
- Dans la table : $P(U < 1.00) \approx 0.8413 \Rightarrow p_c = 0.3174$
- $p_c > \alpha$: On ne rejette pas H_0 car l'observation de cet échantillon ou d'un échantillon encore moins probable sous cette hypothèse est assez vraisemblable (une chance sur trois).

Les autres situations

Nous venons de voir le cas d'un seul échantillon de grande taille.
Mais il existe d'autres situations :

- Petits échantillons (TCL non-adapté)
- Comparaison de deux échantillons indépendants
- Comparaison de deux échantillons appariés
- La v.a. est catégorielle (comparaison de proportions)

Chronologie toujours identique

- Définir la variable aléatoire X.
- Définir la population de référence.
- Choisir les hypothèses à tester.
- Définir la statistique de test sous H_0.
- Définir le risque de 1ère espèce maximum.
- Définir la région critique (RC).
- Application numérique.
- Conclusion.