

Exploitation et traitement des données

COURS 4

Rodolphe Priam, Ingénieur stat & Dr informatique

Partie Test Statistiques

- Rappels du modèle paramétrique
- Test sur un ou deux échantillons de loi de Gauss (normale/gaussienne)
- Test sur un ou deux échantillons de distribution de Bernoulli (binaire)
- Test pour la régressions

Cadre probabilistique: variables aléatoires

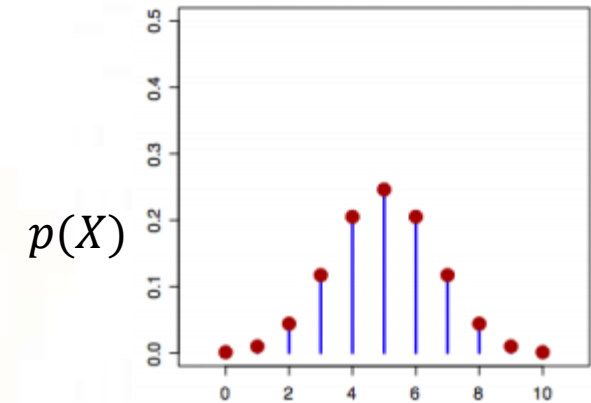
- Informellement, une variable aléatoire (v.a.) X dénote les possible résultats d'un événement tout en exprimant leur probabilités d'occurrence

- Soit discrète (éventuellement nombreux résultats)

Tel que $X \in \{0, 1\}$

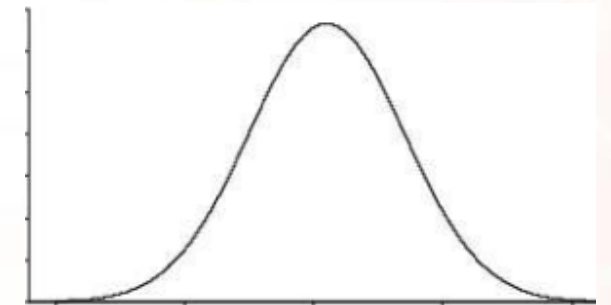
Tel que ou $X \in \{1, 2, \dots, 6\}$

Tel que $X \in \{0, 2, \dots, N\}$ pour N un entier positif



X (a discrete r.v.)

$f(X = x)$

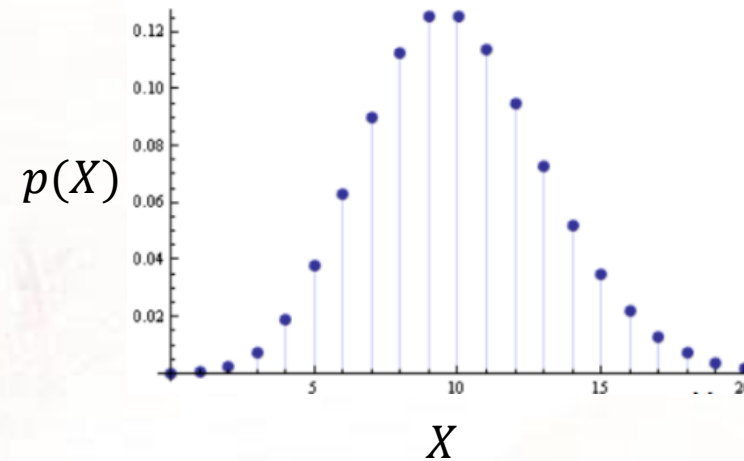


x réalisation d'une v.a. continue

Variable aléatoire discrète: définition


- Pour une v.a. X , $p(x)$ denote $p(X = x)$, la probabilité que $X = x$
- $p(X)$ is est appelée la fonction de probabilité de masse de la v.a. X
 - $p(x)$ ou $p(X = x)$ est la valeur de la fonction pour x

$$\begin{aligned} p(x) &\geq 0 \\ p(x) &\leq 1 \\ \sum_x p(x) &= 1 \end{aligned}$$



Variable aléatoire discrète: exemples

Les données pour les colonnes
cyl, hp, vs, am, gear, carb
sont les réalisations de variables
aléatoires discrètes



rownames	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.9	2.62	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.9	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.32	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.44	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.46	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.57	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.19	20.0	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.15	22.9	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.44	18.3	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.44	18.9	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.07	17.4	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.73	17.6	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.78	18.0	0	0	3	3
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.25	17.98	0	0	3	4
Lincoln Continental	10.4	8	460.0	215	3.0	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.2	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.9	1	1	4	1
Toyota Corona	21.5	4	120.1	97	3.7	2.465	20.01	1	0	3	1
Dodge Challenger	15.5	8	318.0	150	2.76	3.52	16.87	0	0	3	2
AMC Javelin	15.2	8	304.0	150	3.15	3.435	17.3	0	0	3	2
Camaro Z28	13.3	8	350.0	245	3.73	3.84	15.41	0	0	3	4
Pontiac Firebird	19.2	8	400.0	175	3.08	3.845	17.05	0	0	3	2
Fiat X1-9	27.3	4	79.0	66	4.08	1.935	18.9	1	1	4	1
Porsche 914-2	26.0	4	120.3	91	4.43	2.14	16.7	0	1	5	2
Lotus Europa	30.4	4	95.1	113	3.77	1.513	16.9	1	1	5	2
Ford Pantera L	15.8	8	351.0	264	4.22	3.17	14.5	0	1	5	4
Ferrari Dino	19.7	6	145.0	175	3.62	2.77	15.5	0	1	5	6
Maserati Bora	15.0	8	301.0	335	3.54	3.57	14.6	0	1	5	8
Volvo 142E	21.4	4	121.0	109	4.11	2.78	18.6	1	1	4	2

Variable aléatoire discrète: loi de Bernoulli

- Pour X discrète distribuée $B(p)$

$$p(X = 1) = p \text{ et } p(X = 0) = 1 - p$$

$$p(X = 1) + p(X = 0) = 1$$

- Par exemple, la colonne vs est à valeur 0/1

0, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0,

0, 1, 1, 1, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1.

Pour ces seules données, $p = \frac{14}{32} = 0,4375$

Si on considère ces données comme la population!

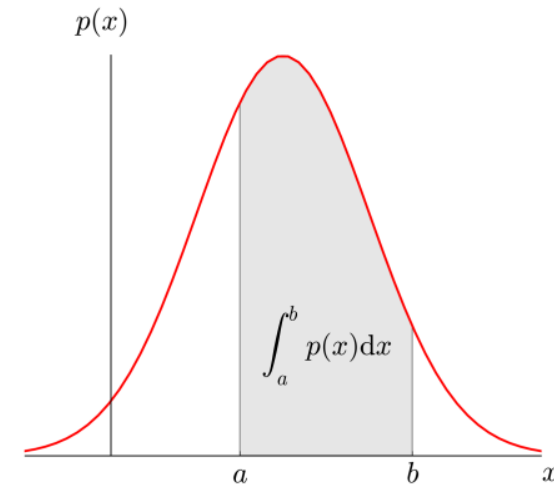
En cas d'un échantillon, p est inconnu: $\hat{p} = 0,4375$

rownames	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.9	2.62	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.9	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.32	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.44	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.46	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.57	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.19	20.0	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.15	22.9	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.44	18.3	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.44	18.9	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.07	17.4	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.73	17.6	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.78	18.0	0	0	3	3
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.25	17.98	0	0	3	4
Lincoln Continental	10.4	8	460.0	215	3.0	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.2	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.9	1	1	4	1
Toyota Corona	21.5	4	120.1	97	3.7	2.465	20.01	1	0	3	1
Dodge Challenger	15.5	8	318.0	150	2.76	3.52	16.87	0	0	3	2
AMC Javelin	15.2	8	304.0	150	3.15	3.435	17.3	0	0	3	2
Camaro Z28	13.3	8	350.0	245	3.73	3.84	15.41	0	0	3	4
Pontiac Firebird	19.2	8	400.0	175	3.08	3.845	17.05	0	0	3	2
Fiat X1-9	27.3	4	79.0	66	4.08	1.935	18.9	1	1	4	1
Porsche 914-2	26.0	4	120.3	91	4.43	2.14	16.7	0	1	5	2
Lotus Europa	30.4	4	95.1	113	3.77	1.513	16.9	1	1	5	2
Ford Pantera L	15.8	8	351.0	264	4.22	3.17	14.5	0	1	5	4
Ferrari Dino	19.7	6	145.0	175	3.62	2.77	15.5	0	1	5	6
Maserati Bora	15.0	8	301.0	335	3.54	3.57	14.6	0	1	5	8
Volvo 142E	21.4	4	121.0	109	4.11	2.78	18.6	1	1	4	2

Variable aléatoire continue: définition

- Pour une v.a. X continue, une probabilité $p(X = x) = p(x)$ a pas de sens ($=0$)
- Pour une v.a..cont., il intervient la prob dans un intervalle $X \in (x, x + \delta x)$
 - $f(x)\delta x$ est la proba que $X \in (x, x + \delta x)$ alors que $\delta x \rightarrow 0$
 - $f(x)$ est la fonction de densité de probabilité à $X = x$

$$\begin{aligned} p(x) &\geq 0 \\ \cancel{p(x)} &\leq 1 \\ \int p(x)dx &= 1 \end{aligned}$$




Variable aléatoire continue: exemples

Les données pour les colonnes

mpg, disp, drat, wt, qsec

sont les réalisations de variables
aléatoires continues



rownames	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.9	2.62	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.9	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.32	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.44	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.46	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.57	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.19	20.0	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.15	22.9	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.44	18.3	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.44	18.9	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.07	17.4	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.73	17.6	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.78	18.0	0	0	3	3
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.25	17.98	0	0	3	4
Lincoln Continental	10.4	8	460.0	215	3.0	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.2	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.9	1	1	4	1
Toyota Corona	21.5	4	120.1	97	3.7	2.465	20.01	1	0	3	1
Dodge Challenger	15.5	8	318.0	150	2.76	3.52	16.87	0	0	3	2
AMC Javelin	15.2	8	304.0	150	3.15	3.435	17.3	0	0	3	2
Camaro Z28	13.3	8	350.0	245	3.73	3.84	15.41	0	0	3	4
Pontiac Firebird	19.2	8	400.0	175	3.08	3.845	17.05	0	0	3	2
Fiat X1-9	27.3	4	79.0	66	4.08	1.935	18.9	1	1	4	1
Porsche 914-2	26.0	4	120.3	91	4.43	2.14	16.7	0	1	5	2
Lotus Europa	30.4	4	95.1	113	3.77	1.513	16.9	1	1	5	2
Ford Pantera L	15.8	8	351.0	264	4.22	3.17	14.5	0	1	5	4
Ferrari Dino	19.7	6	145.0	175	3.62	2.77	15.5	0	1	5	6
Maserati Bora	15.0	8	301.0	335	3.54	3.57	14.6	0	1	5	8
Volvo 142E	21.4	4	121.0	109	4.11	2.78	18.6	1	1	4	2

Variable aléatoire continue: loi Gaussienne/Normale⁹

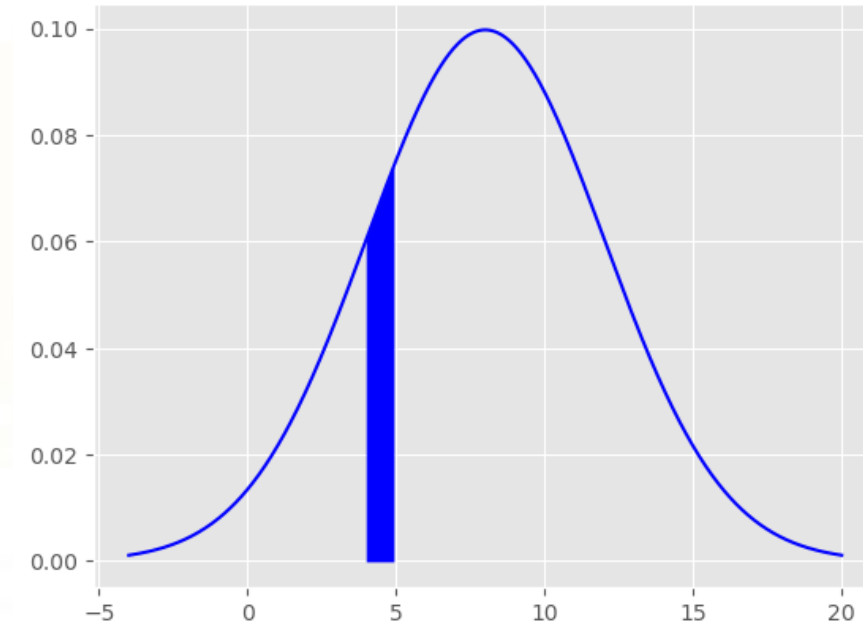
- $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-0,5\left(\frac{x-m}{\sigma}\right)^2}$ symétrique, de moyenne m et écart-type σ .

- Pour rappel, il y a une forte chance/probabilité que les observations se trouvent proches de m

- Exemple $X \sim N(8,4)$

- On peut calculer $P(X \in [4,5]) = \int_4^5 f(x)dx$
- Graphiquement comme une aire sous la fonction
- Algébriquement car:

$$\begin{aligned} P(X \in [4,5]) &= P(4 \leq X \leq 5) = P\left(\frac{4-8}{4} \leq \frac{X-8}{4} \leq \frac{5-8}{4}\right) \\ &= P(-1 \leq Z \leq -0.75) \text{ avec } Z \sim N(0,1) \text{ standard tabulée !} \end{aligned}$$



$P(X \in [4,5])$

Variable aléatoire continue: autres lois

- Loi de Student symétrique généralise la loi normale centrée réduite $N(0,1)$
 - Utile pour les intervalles de confiance et pour les tests de moyenne car:

$$T = \frac{Z}{\sqrt{U/k}}$$

T loi de student à k degrés de liberté

Z une variable aléatoire de loi normale centrée et réduite

U une variable indépendante de Z et distribuée suivant la loi du χ^2 à k degrés de liberté

- Loi du χ^2 asymétrique converge vers une loi normale
 - Utile pour les intervalle de confiance des variances et tests de distribution (hors sujet ici)

$$U = \sum_{i=1}^k X_i^2$$

U loi du χ^2 à k degrés de liberté

X_i variables aléatoires réelles i.i.d. de loi $N(0,1)$, normales centrées-réduites

Remarque sur la notation

- $p(X)$ denote généralement la distribution (PMF/PDF) d'une v.a. X
- $f(.)$ est parfois noté $p(.)$ également suivant le domaine (ex: machine learning)

$$x \sim p(X)$$

$$x \sim f(X)$$

- $p(X = x)$ ou $p_X(x)$ ou $p(x)$ dénote la proba ou densité de proba en x
- Lorsque $p(.)$ prend une forme spécifique les statistiques $g()$ sont différentes
 - Cela explique pourquoi les intervalles de confiances et tests sont présenté pour des cas particuliers, puisque les distributions vont être modifiées d'un cas à l'autre
 - Une fois compris l'approche générale, les tests se ressemblent, même si la loi change.

- Population inconnue dispose certaines caractéristiques modélisables
 - Distribution connue telle Gaussienne (moyenne, variance) ou Bernoulli (proportion)
 - Paramètres (inconnus) telles que moyenne=?, variance=? ou proportion=?
 - Formellement une variable aléatoire mène à la réalisation d'une observation

soit	$X_i \sim \text{Gaussienne}(\text{moyenne}, \text{variance})$	\Rightarrow	x_i dans R
soit	$X_i \sim \text{Bernoulli}(\text{proportion})$	\Rightarrow	x_i dans $\{0,1\}$

- Echantillon connu
 - ensemble des observations disponibles $S = (x_1, \dots, x_n)$
 - Les observations x_1, \dots, x_n sont des réalisations i.i.d. des v.a. X_1, \dots, X_n :

$$X_1 \Rightarrow x_1, X_2 \Rightarrow x_2, \dots, X_n \Rightarrow x_n.$$
 - Contexte d'une expérience répétée qui génère chaque x_i de l'échantillon.
- Objectif: tirer de conclusions valides à partir de l'échantillon!

- Population inconnue

- Distribution connue Les variables aléatoires X_i sont de loi connue/supposée
- Paramètres (inconnus) La loi des X_i est connue mais pas ses paramètres
- Exemple

soit $X_i \sim \text{Gaussienne}(\text{moyenne}, \text{variance})$ avec moyenne=? variance=?

soit $X_i \sim \text{Bernoulli}(\text{proportion})$ avec proportion $p=?$

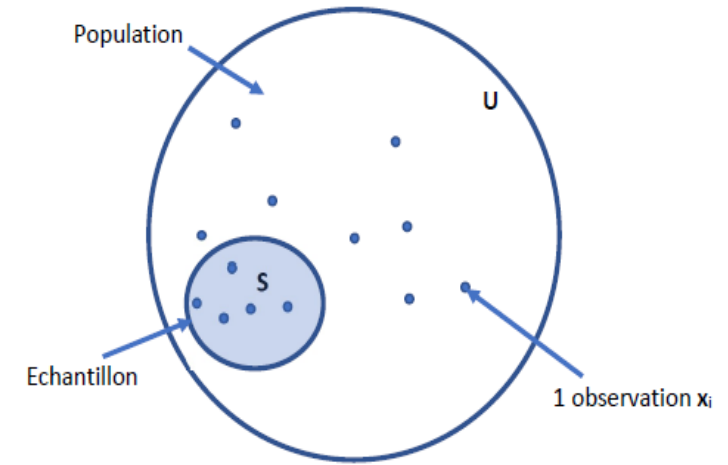
- Echantillon connu

- Ce sont les données telles que chaque x_i est une réalisation d'une v.a. X_i

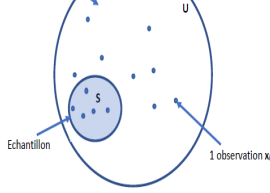
- Conséquence

- La statistique g a une loi connue (parfois seulement approchée car nonlinéaire)
- $\theta_n = g(X_1, \dots, X_n)$ est une variable aléatoire, $\hat{\theta}_n = g(x_1, \dots, x_n)$ en est une réalisation

- Hypothèse: l'échantillon de données provient d'une population U
 - L'échantillon est de taille finie n (petit ou large)
 - La population est de taille infinie (ou très très large)
 - Les échantillons sont des échantillons aléatoires, dans le sens que les individus sélectionnés dans U ont eu la même chance que tous les autres dans U.
 - L'échantillon S est un échantillon représentatif il représente une image réaliste de la population.



- L'inférence statistique est le domaine des statistiques qui permet d'établir des faits concernant la population à partir des résultats obtenus en étudiant seulement l'échantillon $S = (x_1, \dots, x_n)$
 - Exemple: quelle est la moyenne ou la variance d'une caractéristique dans U ?
 - Exemple: quelle est la proportion de succès d'un événement dans U ?

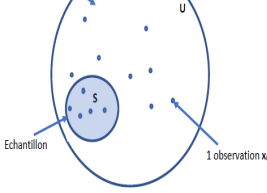


- Hypothèse: l'échantillon de données provient d'une population U
- Population inconnue dispose certaines caractéristiques modélisables
- Echantillon connu ensemble des observations disponibles $S = (x_1, \dots, x_n)$

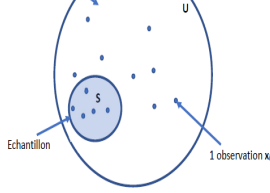
- Exemple de question posée

- **Estimer** une caractéristique de la pop. U
 - Quelle est la puissance moyenne des voitures dans l'ensemble de la population U (inconnue)
 - Quelle est la proportion de voit. ayant 5 vitesses dans l'ensemble de la population U (inconnue)
- **Tester** une hypothèse sur la caractéristique
 - La puissance moyenne est-elle 120 chevaux ?
 - La proportion de 5 vitesses est-elle 70% ?

rownames	mpg	cyl	displacement	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.9	2.62	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.9	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.32	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.44	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.46	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.57	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.19	20.0	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.15	22.9	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.44	18.3	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.44	18.9	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.07	17.4	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.73	17.6	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.78	18.0	0	0	3	3
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.25	17.98	0	0	3	4
Lincoln Continental	10.4	8	460.0	215	3.0	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.2	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.9	1	1	4	1
Toyota Corona	21.5	4	120.1	97	3.7	2.465	20.01	1	0	3	1
Dodge Challenger	15.5	8	318.0	150	2.76	3.52	16.87	0	0	3	2
AMC Javelin	15.2	8	304.0	150	3.15	3.435	17.3	0	0	3	2
Camaro Z28	13.3	8	350.0	245	3.73	3.84	15.41	0	0	3	4
Pontiac Firebird	19.2	8	400.0	175	3.08	3.845	17.05	0	0	3	2
Fiat X1-9	27.3	4	79.0	66	4.08	1.935	18.9	1	1	4	1
Porsche 914-2	26.0	4	120.3	91	4.43	2.14	16.7	0	1	5	2
Lotus Europa	30.4	4	95.1	113	3.77	1.513	16.9	1	1	5	2
Ford Pantera L	15.8	8	351.0	264	4.22	3.17	14.5	0	1	5	4
Ferrari Dino	19.7	6	145.0	175	3.62	2.77	15.5	0	1	5	6
Maserati Bora	15.0	8	301.0	335	3.54	3.57	14.6	0	1	5	8
Volvo 142E	21.4	4	121.0	109	4.11	2.78	18.6	1	1	4	2



- Hypothèse: l'échantillon de données provient d'une population U
- Population inconnue dispose certaines caractéristiques modélisables
- Echantillon connu ensemble des observations disponibles $S = (x_1, \dots, x_n)$
- Solution:
 - Calcul de la statistique $\hat{\theta}_n = g(x_1, \dots, x_n)$
 - La moyenne inconnue de la population est remplacée par la moyenne de l'échantillon !
 - La variance inconnue de la population est remplacée par la variance de l'échantillon !
 - La proportion inconnue de la population est remplacée par la proportion de l'échantillon !
 - Calcul d'un intervalle $I_n = [\min, \max]$ tel que θ est dans I_n
 - l'hypothèse sur la population U et la loi de X_i conduit à trouver la loi de $\theta_n = g(X_1, \dots, X_n)$
 - Il est déduit un encadrement du vrai paramètre θ inconnu à partir d'hypothèses sur θ_n



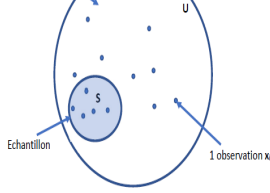
- Données : observations x_1, \dots, x_n
 - Modèle : variables aléatoires X_1, \dots, X_n i.i.d.
(distribution connue, paramètres inconnus)
-

- Une statistique d'échantillonnage ou statistique est quantité calculée à partir des variables aléatoires, il s'agit d'une fonction des X_i :

$$\theta_n = g(X_1, \dots, X_n) \quad \text{Estimateur}$$

- A cette statistique correspond la valeur correspondant à l'échantillon observée, pour l'ensemble des observations disponibles:

$$\hat{\theta}_n = g(x_1, \dots, x_n) \quad \text{Estimation}$$

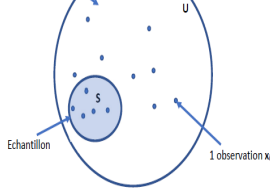


- Données : observations x_1, \dots, x_n
 - Modèle : variables aléatoires X_1, \dots, X_n i.i.d.
(distribution connue, paramètres inconnus)
-

- Une statistique d'échantillonnage ou statistique est quantité calculée à partir des variables aléatoires, il s'agit d'une fonction des X_i :

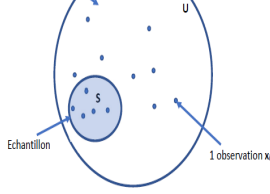
$$\theta_n = g(X_1, \dots, X_n) \quad \text{Estimateur de } \theta \text{ inconnu.}$$

- Biais et variance de l'estimateur de θ sont des propriétés essentielles
 - Biais : $B(\theta_n) = E[\theta_n] - \theta$
 - Variance : $V(\theta_n) = E(\theta_n^2) - E(\theta_n)^2$



- Données : observations x_1, \dots, x_n
 - Modèle : variables aléatoires X_1, \dots, X_n i.i.d.
(distribution connue, paramètres inconnus)
-

- $X_i \sim N(m, \sigma)$ donc $E[X_i] = m$ et $V[X_i] = \sigma^2$
- Estimateur \bar{X}_n de la moyenne m
 - $\bar{X}_n = \frac{X_1 + X_2 + \dots + X_n}{n}$ (et estimation \bar{x}_n) .
 - $\bar{X}_n \sim N(m, \frac{\sigma}{\sqrt{n}})$ donc $E[\bar{X}_n] = m$ et $V[\bar{X}_n] = \frac{\sigma^2}{n}$ (et \bar{x}_n réalisation de \bar{X}_n)
 - Cette statistique est effectivement de loi connue dépendant de celle des X_i .



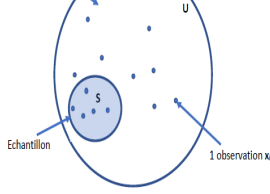
- Données : observations x_1, \dots, x_n
 - Modèle : variables aléatoires X_1, \dots, X_n i.i.d.
(distribution connue, paramètres inconnus)
-

- Preuve que si $X_i \sim N(m, \sigma)$ alors $\bar{X}_n \sim N(m, \frac{\sigma}{\sqrt{n}})$

- $$E[\bar{X}_n] = E\left[\frac{X_1 + X_2 + \dots + X_n}{n}\right] = \frac{E[X_1] + E[X_2] + \dots + E[X_n]}{n} = \frac{n m}{n} = m$$

- $$V[\bar{X}_n] = \frac{V[X_1] + V[X_2] + \dots + V[X_n]}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

- Une combinaison de lois normales reste une loi normale (admis).



• Données : observations x_1, \dots, x_n Modèle : v.a. X_1, \dots, X_n i.i.d.

• Si $E(\bar{X}_n) = \mu$ alors \bar{X}_n est non biaisé, $E(\bar{X}_n) = \mu$, mais S_n avec $\frac{1}{n}$ est biaisé.

1

$$\begin{aligned}
 E[S^2] &= E \left[\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \right] = E \left[\frac{1}{n} \sum_{i=1}^n \left((X_i - \mu) - (\bar{X} - \mu) \right)^2 \right] \\
 &= E \left[\frac{1}{n} \sum_{i=1}^n \left((X_i - \mu)^2 - 2(\bar{X} - \mu)(X_i - \mu) + (\bar{X} - \mu)^2 \right) \right] \\
 &= E \left[\frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 - \frac{2}{n} (\bar{X} - \mu) \sum_{i=1}^n (X_i - \mu) + \frac{1}{n} (\bar{X} - \mu)^2 \sum_{i=1}^n 1 \right] \\
 &= E \left[\frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 - \frac{2}{n} (\bar{X} - \mu) \sum_{i=1}^n (X_i - \mu) + \frac{1}{n} (\bar{X} - \mu)^2 \cdot n \right] \\
 &= E \left[\frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 - \frac{2}{n} (\bar{X} - \mu) \sum_{i=1}^n (X_i - \mu) + (\bar{X} - \mu)^2 \right]
 \end{aligned}$$

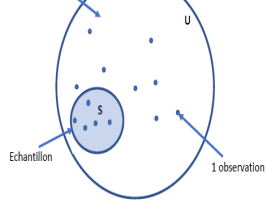
3

$$\begin{aligned}
 E[S^2] &= E \left[\frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 - \frac{2}{n} (\bar{X} - \mu) \sum_{i=1}^n (X_i - \mu) + (\bar{X} - \mu)^2 \right] \\
 &= E \left[\frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 - \frac{2}{n} (\bar{X} - \mu) \cdot n \cdot (\bar{X} - \mu) + (\bar{X} - \mu)^2 \right] \\
 &= E \left[\frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 - 2(\bar{X} - \mu)^2 + (\bar{X} - \mu)^2 \right] \\
 &= E \left[\frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 - (\bar{X} - \mu)^2 \right] \\
 &= E \left[\frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 \right] - E \left[(\bar{X} - \mu)^2 \right] \\
 &= \sigma^2 - E \left[(\bar{X} - \mu)^2 \right] = \left(1 - \frac{1}{n} \right) \sigma^2 < \sigma^2
 \end{aligned}$$

Source: Wikipédia

2

$$\bar{X} - \mu = \frac{1}{n} \sum_{i=1}^n X_i - \mu = \frac{1}{n} \sum_{i=1}^n X_i - \frac{1}{n} \sum_{i=1}^n \mu = \frac{1}{n} \sum_{i=1}^n (X_i - \mu).$$

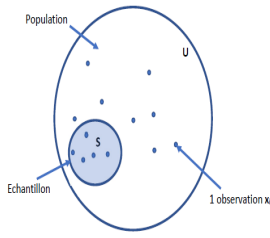


- Données : observations x_1, \dots, x_n Modèle : v.a. X_1, \dots, X_n i.i.d.
- Si $E(\bar{X}_n) = \mu$ alors \bar{X}_n et S_n avec $\frac{1}{n-1}$ sont non biaisés.
- S_n avec $\frac{1}{n-1}$ est non biaisé (version préférée dans la suite)

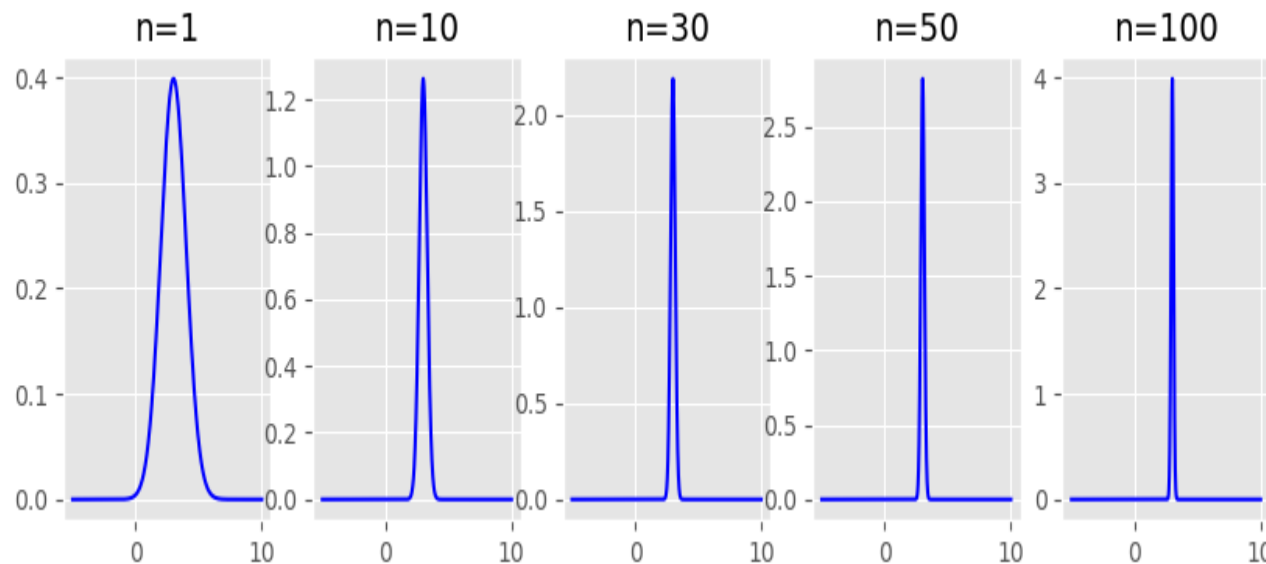
$$s_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

- On vérifie facilement le non biais d'après le résultat précédent:

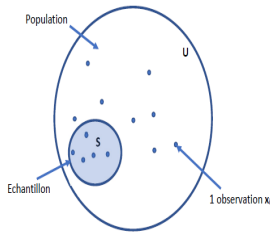
$$\begin{aligned} E[S^2] &= E\left[\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2\right] = \frac{n}{n-1} E\left[\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2\right] \\ &= \frac{n}{n-1} \left(1 - \frac{1}{n}\right) \sigma^2 = \sigma^2, \end{aligned}$$



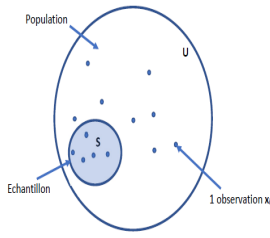
- Données : observations x_1, \dots, x_n
- Modèle : variables aléatoires X_1, \dots, X_n i.i.d.
- Si $X_i \sim N(m, \sigma)$ alors $\bar{X}_n \sim N(m, \frac{\sigma}{\sqrt{n}})$.
- Graphiquement par exemple avec $m=3$ et $\sigma=1$.



Lorsque la taille de l'échantillon augmente les valeurs des moyennes des échantillons se concentrent autour de la vraie moyenne si bien que pour n très grand on obtient finalement $\bar{X}_n = m$.

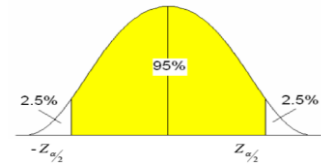


- Données : observations x_1, \dots, x_n
- Modèle : variables aléatoires X_1, \dots, X_n i.i.d.
- $X_i \sim B(p)$ donc $E[X_i] = p$ et $V[X_i] = p(1-p)$
- Estimateur F_n de la proportion p (et estimation f_n).
 - $$F_n = \frac{X_1 + X_2 + \dots + X_n}{n}$$
 - $F_n \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$ asymptot. pour $np, nq \geq 5$ et $n > 30$ (et f_n réalisation de F_n).
 - $E[F_n] = p$ et $V[F_n] = \frac{p(1-p)}{n}$
 - Cette statistique est effectivement de loi connue dépendant de celle des X_i .



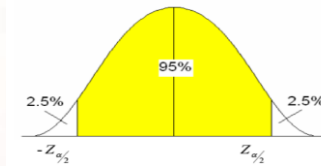
- Données : observations x_1, \dots, x_n
- Modèle : variables aléatoires X_1, \dots, X_n i.i.d.
- Intervalle de confiance pour 1 moyenne obtenu car $\bar{X}_n \sim N\left(m, \frac{\sigma}{\sqrt{n}}\right)$

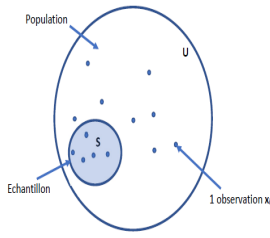
$$\left[\bar{X} - Z_{\alpha} \frac{\sigma}{\sqrt{n}}; \bar{X} + Z_{\alpha} \frac{\sigma}{\sqrt{n}} \right] \text{ avec probabilité } 1-\alpha.$$



- Intervalle de confiance pour 1 proportion obtenue car $F_n \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$

$$\left[F - Z_{\alpha} \sqrt{\frac{pq}{n}}; F + Z_{\alpha} \sqrt{\frac{pq}{n}} \right] \text{ avec probabilité } 1-\alpha.$$





- Données : observations x_1, \dots, x_n
- Modèle : variables aléatoires X_1, \dots, X_n i.i.d.
- Justification des intervalles vu précédemment
- Justification de la standardisation pour $\bar{X}_n \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$:

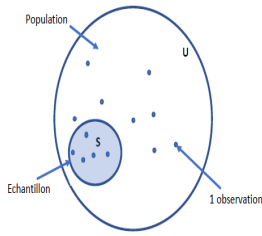
$$Z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \text{ statistique standardisée.}$$

$$E(Z) = E\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}\right) = \frac{1}{\sigma/\sqrt{n}} E(\bar{X}_n - \mu) = \frac{0}{\sigma/\sqrt{n}} = 0$$

$$V(Z) = V\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}\right) = \frac{1}{\sigma^2/n} V(\bar{X}_n - \mu) = \frac{1}{\sigma^2/n} V(\bar{X}_n) = \frac{\sigma^2/n}{\sigma^2/n} = 1$$

D'où,

$$Z \sim N(0,1), \text{ loi normale centré-réduite.}$$



- Données : observations x_1, \dots, x_n
- Modèle : variables aléatoires X_1, \dots, X_n i.i.d.
- Exemple d'intervalle pour p car $F_n \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$

$$\left[F - Z_\alpha \sqrt{\frac{pq}{n}}; F + Z_\alpha \sqrt{\frac{pq}{n}} \right]$$

Reprenons la colonne **vs**

0, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1,

0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0,

0, 0, 0, 1, 0, 1, 0, 0, 0, 1

On calcule: $\hat{p} = f_n = 0,4375$
 Donc la vraie proportion p est telle
 que $p \in [I_{\min}, I_{\max}]$ avec proba 0,95:

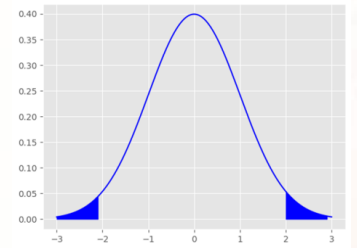
$$I_{\min} = f_n - 1.96 \sqrt{\frac{f_n(1-f_n)}{n}} = 0.27$$

$$I_{\max} = f_n + 1.96 \sqrt{\frac{f_n(1-f_n)}{n}} = 0.61$$

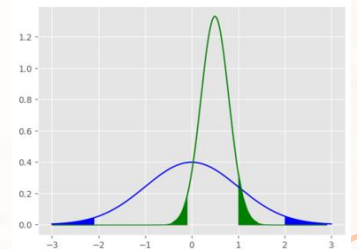
rownames	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.9	2.62	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.9	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.32	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet											
Sportabout	18.7	8	360.0	175	3.15	3.44	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.46	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.57	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.19	20.0	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.15	22.9	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.44	18.3	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.44	18.9	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.07	17.4	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.73	17.6	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.78	18.0	0	0	3	3
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.25	17.98	0	0	3	4
Lincoln Continental	10.4	8	460.0	215	3.0	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.2	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.9	1	1	4	1
Toyota Corona	21.5	4	120.1	97	3.7	2.465	20.01	1	0	3	1
Dodge Challenger	15.5	8	318.0	150	2.76	3.52	16.87	0	0	3	2
AMC Javelin	15.2	8	304.0	150	3.15	3.435	17.3	0	0	3	2
Camaro Z28	13.3	8	350.0	245	3.73	3.84	15.41	0	0	3	4
Pontiac Firebird	19.2	8	400.0	175	3.08	3.845	17.05	0	0	3	2
Fiat X1-9	27.3	4	79.0	66	4.08	1.935	18.9	1	1	4	1
Porsche 914-2	26.0	4	120.3	91	4.43	2.14	16.7	0	1	5	2
Lotus Europa	30.4	4	95.1	113	3.77	1.513	16.9	1	1	5	2
Ford Pantera L	15.8	8	351.0	264	4.22	3.17	14.5	0	1	5	4
Ferrari Dino	19.7	6	145.0	175	3.62	2.77	15.5	0	1	5	6
Maserati Bora	15.0	8	301.0	335	3.54	3.57	14.6	0	1	5	8
Volvo 142E	21.4	4	121.0	109	4.11	2.78	18.6	1	1	4	2

- Données : observations x_1, \dots, x_n (et y_1, \dots, y_n)
- Modèle : variables aléatoires X_1, \dots, X_{n1} i.i.d. (et Y_1, \dots, Y_{n2} i.i.d.)
- Classiquement pour un ou deux échantillons

- Cas 1 échantillon: Pour $X_i \sim N(m, \sigma)$ pour continu (sinon $B(p)$ si proportion p)
 - on veut tester /décider si $m=\mu$ pour μ un nombre donné
 - Exemple:
 - Quid du test pour une proportion $p = \mu$?



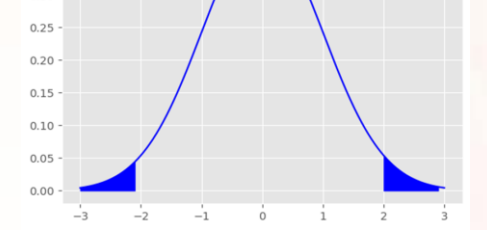
- Cas 2 échantillon: Pour $X_i \sim N(m_x, \sigma_x)$ et $Y_i \sim N(m_y, \sigma_y)$ ou sinon $B(p_x), B(p_y)$
 - on veut tester/décider si $m_x=m_y$ pour μ un nombre donné
 - Exemple:
 - Quid du test pour deux proportions à comparer $p_x = p_y$?



- Données : observations x_1, \dots, x_n (et y_1, \dots, y_n)
- Modèle : variables aléatoires X_1, \dots, X_n i.i.d. (et Y_1, \dots, Y_n i.i.d.)
- On veut tester une hypothèse sur la population
 - Hypothèse H_0 contre hypothèse H_1
 - Exemple $m = \mu$ pour μ une valeur numérique (échantillon des x_i)
 - Exemple $p = \mu$ pour μ une valeur numérique (échantillon des x_i)
 - Exemple $m_1 = m_2$ (deux échantillons, celui des x_i et celui des y_i)
 - Exemple $p_1 = p_2$ (deux échantillons, celui des x_i et celui des y_i)
 - Comment décider quelle hypothèse choisir: Peut-on accepter H_0 ?
 - Comme le test est conservatif, on préfère dire, « ne pas rejeter H_0 » !
 - L'approche se base sur des intervalles pour des lois standardisées

- Données : observations x_1, \dots, x_n
- Modèle : variables aléatoires X_1, \dots, X_n i.i.d.
- Si $X_i \sim N(m, \sigma)$ alors $\bar{X}_n \sim N(m, \sigma/\sqrt{n})$.
- Si H_0 suppose $m=\mu$ alors $\bar{X}_n \sim N(\mu, \sigma/\sqrt{n})$ et \bar{x}_n réalisation de \bar{X}_n .
 - Donc on vérifie que la valeur de \bar{x}_n est assez probable pour cette distribution.
 - On calcule un intervalle I_α dans lequel \bar{x}_n a $\alpha=95\%$ de chance de se trouver.
 - Si la moyenne empirique se trouve dans l'intervalle on rejette pas H_0 .
- Deux situations, soit σ connu soit σ inconnu estimé suivant le cas

σ connu	σ inconnu estimé par S_n
$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$	$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim \mathcal{T}_{n-1}$
loi normale standard $N(0,1)$	loi de Student à $n-1$ ddl



- Données : observations x_1, \dots, x_n
- Modèle : variables aléatoires X_1, \dots, X_n i.i.d.

σ connu donne une loi normale standard $\mathcal{N}(0,1)$

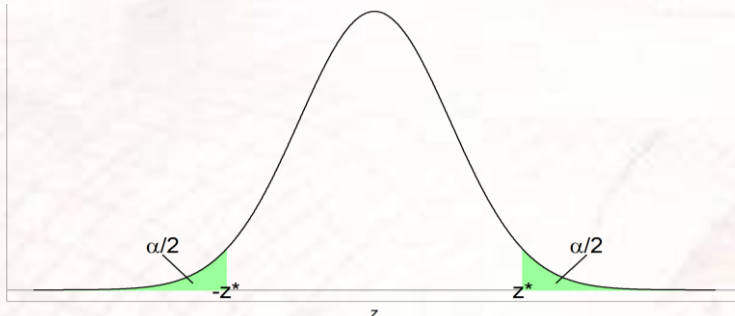
$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

- Calcul de $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$
- Pour test à $1-\alpha$ et $\alpha=95\%$ si $z \in [-z_\alpha, +z_\alpha]$, H_0 du t-test non rejeté

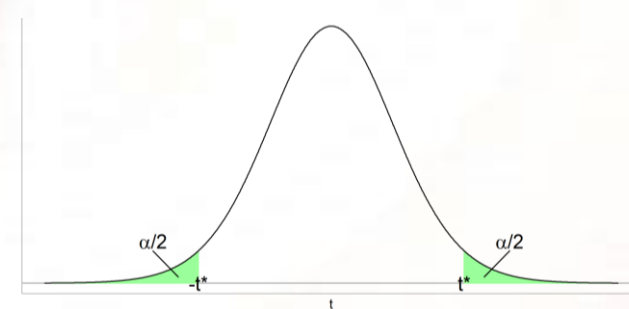
σ inconnu estimé par S_n donne une loi de Student à $n-1$ ddl

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim \mathcal{T}_{n-1}$$

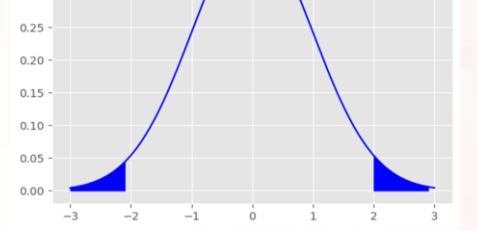
- Calcul de $t = \frac{\bar{x} - \mu}{s_n/\sqrt{n}}$
- Pour test à $1-\alpha$ et $\alpha=95\%$ si $t \in [-t_{n-1,\alpha}, +t_{n-1,\alpha}]$, H_0 du t-test non rejeté



Pour $\alpha=95\%$, on peut calculer $z_\alpha = 1,96$



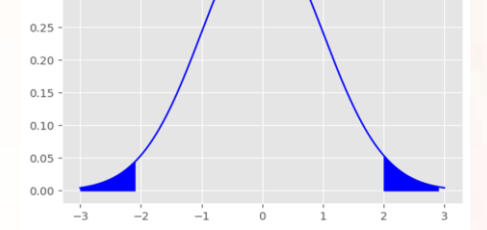
Pour $\alpha=95\%$, on peut calculer $t_{n-1,\alpha}$ pour valeur n



- Données : observations x_1, \dots, x_n
- Modèle : variables aléatoires X_1, \dots, X_n i.i.d.
- Si $X_i \sim B(p)$ alors $F_n \sim N(p, \sqrt{p(1-p)/n})$ asymptotiquement.
- Si H_0 suppose $p = p_0$ alors $F_n \sim N(p_0, \sqrt{p_0(1-p_0)/n})$ et f_n réalisation de F_n
 - Donc on vérifie que la valeur de f_n est assez probable pour cette distribution.
 - On calcule un intervalle I_α dans lequel f_n a $\alpha=95\%$ de chance de se trouver.
 - Si la moyenne empirique se trouve dans l'intervalle on rejette pas H_0 .
- Un seul cas pour une proportion (au contraire de la moyenne)

loi normale standard $N(0,1)$ asymptotique pour $np, n(1-p) \geq 5$ et $n > 30$

$$Z = \frac{F_n - p_0}{\sqrt{p_0(1-p_0)}/\sqrt{n}} \sim \mathcal{N}(0,1)$$



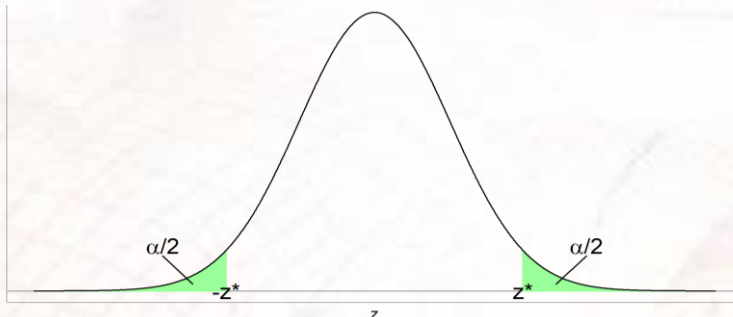
- Données : observations x_1, \dots, x_n
- Modèle : variables aléatoires X_1, \dots, X_n i.i.d.

$$Z = \frac{F_n - p_0}{\sqrt{p_0(1 - p_0)/\sqrt{n}}} \sim \mathcal{N}(0,1)$$

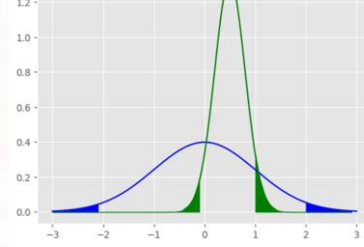
- Calcul de $z = \frac{f_n - p_0}{\sqrt{p_0(1 - p_0)/\sqrt{n}}}$
- Pour test à $1-\alpha$ et $\alpha=95\%$ si $z \in [-z_\alpha, +z_\alpha]$, H_0 non rejeté



Asymptotique, valide
pour $np, n(1-p) \geq 5$ et $n > 30$

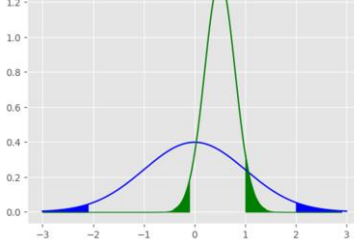


Pour $\alpha=95\%$, on peut calculer $z_\alpha = 1,96$



- Données : observations x_1, \dots, x_n et y_1, \dots, y_n
 - Modèle : variables aléatoires X_1, \dots, X_{n_1} i.i.d. et Y_1, \dots, Y_{n_2} i.i.d.
-
- Si $X_i \sim N(m_1, \sigma_1)$ alors $\bar{X}_n \sim N(m_1, \frac{\sigma_1}{\sqrt{n_1}})$.
 - Si $Y_i \sim N(m_2, \sigma_2)$ alors $\bar{Y}_n \sim N(m_2, \frac{\sigma_2}{\sqrt{n_2}})$.
 - Deux moyennes d'échantillons au lieu d'une seule moyenne ici !
-
- Assez proche du t-test précédent (calcul d'une stat), mais quatre cas !
 - variances connues et égales
 - variances connues et égales
 - variances inconnues et inégales
 - variances inconnues et inégales

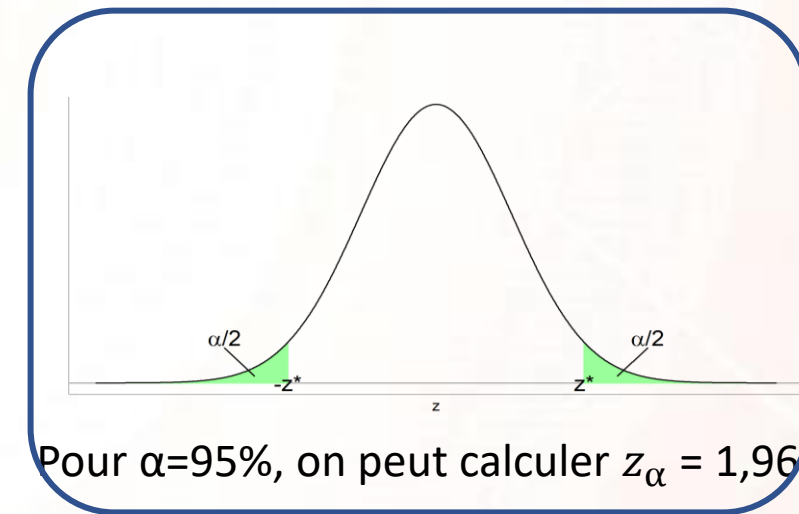
- Données : observations x_1, \dots, x_n et y_1, \dots, y_n
- Modèle : variables aléatoires X_1, \dots, X_{n_1} i.i.d. et Y_1, \dots, Y_{n_2} i.i.d.
- On a $X_i \sim N(m_1, \sigma_1)$, $\bar{X}_n \sim N(m_1, \frac{\sigma_1}{\sqrt{n_1}})$, et $Y_i \sim N(m_2, \sigma_2)$, $\bar{Y}_n \sim N(m_2, \frac{\sigma_2}{\sqrt{n_2}})$.



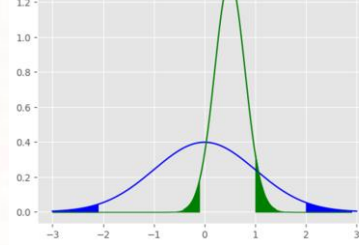
Variances connues et égales ($\sigma_1 = \sigma_2$)

$$Z = \frac{(\bar{X}_n - \bar{Y}_n) - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim \mathcal{N}(0,1) \text{ avec } \mu_1 - \mu_2 = 0$$

Calcul de $z = \frac{\bar{x}_n - \bar{y}_n}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$, alors pour test à $1-\alpha$ et $\alpha=95\%$ si $z \in [-z_\alpha, +z_\alpha]$, H_0 non rejeté



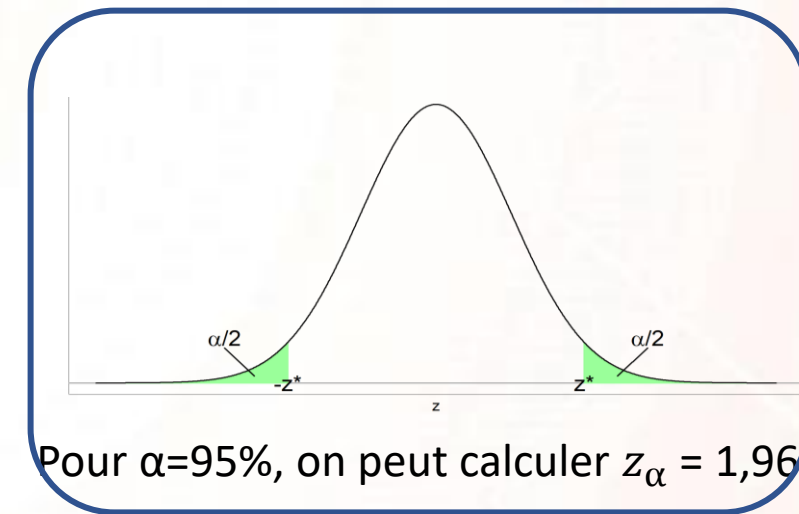
- Données : observations x_1, \dots, x_n et y_1, \dots, y_n
- Modèle : variables aléatoires X_1, \dots, X_{n_1} i.i.d. et Y_1, \dots, Y_{n_2} i.i.d.
- On a $X_i \sim N(m_1, \sigma_1)$, $\bar{X}_n \sim N(m_1, \frac{\sigma_1}{\sqrt{n_1}})$, et $Y_i \sim N(m_2, \sigma_2)$, $\bar{Y}_n \sim N(m_2, \frac{\sigma_2}{\sqrt{n_2}})$.



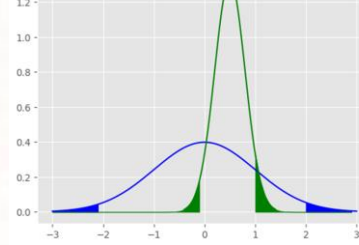
Variances connues et inégales ($\sigma_1 \neq \sigma_2$)

$$Z = \frac{(\bar{X}_n - \bar{Y}_n) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0,1) \text{ avec } \mu_1 - \mu_2 = 0$$

Calcul de $z = \frac{\bar{x}_n - \bar{y}_n}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$, alors pour test à $1-\alpha$ et $\alpha=95\%$ si $z \in [-z_\alpha, +z_\alpha]$, H_0 non rejeté



- Données : observations x_1, \dots, x_n et y_1, \dots, y_n
- Modèle : variables aléatoires X_1, \dots, X_{n_1} i.i.d. et Y_1, \dots, Y_{n_2} i.i.d.
- On a $X_i \sim N(m_1, \sigma_1)$, $\bar{X}_n \sim N(m_1, \frac{\sigma_1}{\sqrt{n_1}})$, et $Y_i \sim N(m_2, \sigma_2)$, $\bar{Y}_n \sim N(m_2, \frac{\sigma_2}{\sqrt{n_2}})$.

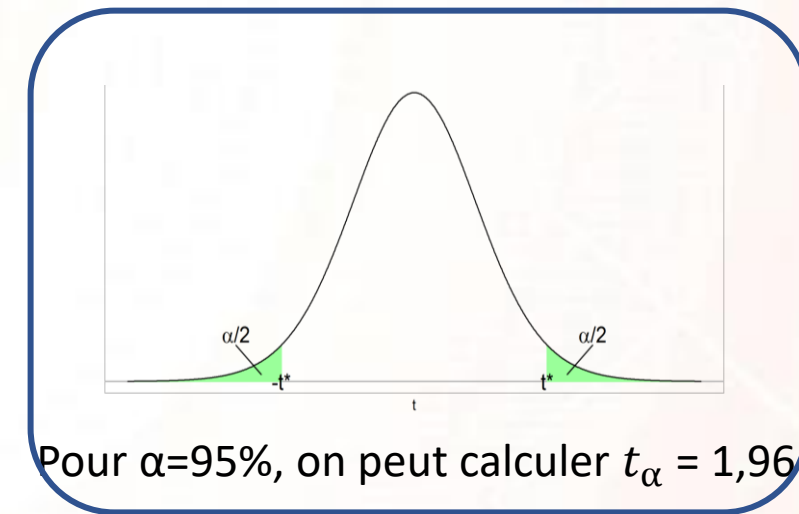


Variances inconnues et égales ($\sigma_1 = \sigma_2$)

$$Z = \frac{(\bar{X}_n - \bar{Y}_n) - (\mu_1 - \mu_2)}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim \mathcal{T}_{n_1 + n_2 - 2} \text{ avec } \mu_1 - \mu_2 = 0$$

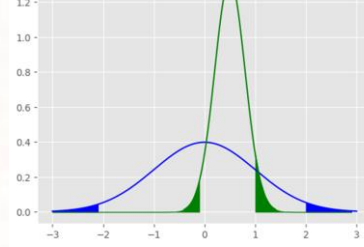
$$\text{Et } S^2 = \frac{(n_1 - 1)S_{n_1}^2 + (n_2 - 1)S_{n_2}^2}{n_1 + n_2}$$

Calcul de $z = \frac{\bar{x}_n - \bar{y}_n}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$, alors pour test à $1-\alpha$ et $\alpha=95\%$ si $z \in [-t_\alpha, +t_\alpha]$, H_0 non rejeté



- Cas variances inconnues et inégales non présenté ici.

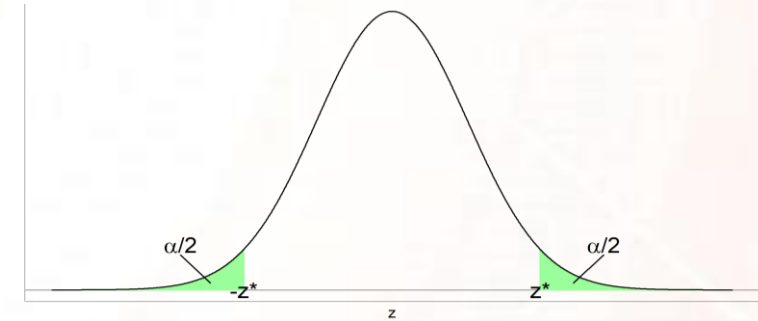
- Données : observations x_1, \dots, x_n et y_1, \dots, y_n
- Modèle : variables aléatoires X_1, \dots, X_{n_1} i.i.d. et Y_1, \dots, Y_{n_2} i.i.d.



- Si $X_i \sim B(p_1)$ alors $F_{n_1} \sim N(p_1, \sqrt{p_1(1-p_1)/n_1})$
- Si $Y_i \sim B(p_2)$ alors $F_{n_2} \sim N(p_2, \sqrt{p_2(1-p_2)/n_2})$

$$Z = \frac{F_{n_1} - F_{n_2}}{\sqrt{F(1-F)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim \mathcal{N}(0,1) \text{ avec } F = \frac{n_1 F_{n_1} - n_2 F_{n_2}}{n_1 + n_2}$$

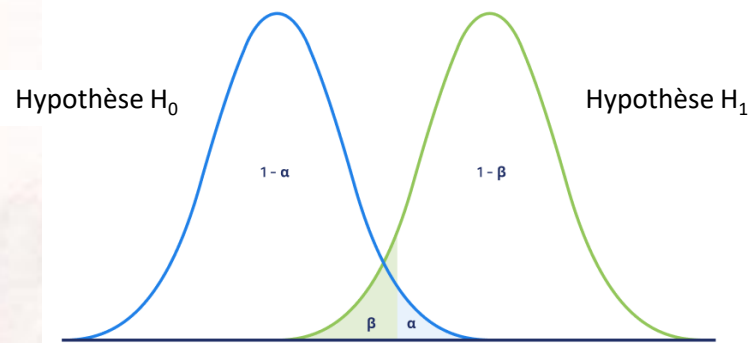
- Calcul de $z = \frac{f_{n_1} - f_{n_2}}{\sqrt{f(1-f)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ avec $f = \frac{n_1 f_{n_1} - n_2 f_{n_2}}{n_1 + n_2}$
- Pour test à $1-\alpha$ et $\alpha=95\%$ si $z \in [-z_\alpha, +z_\alpha]$, H_0 non rejeté



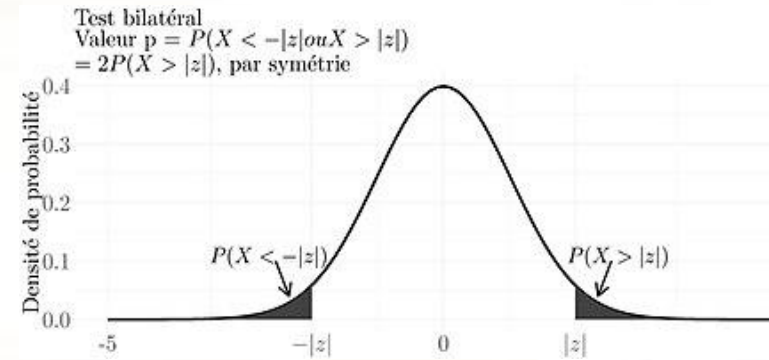
Pour $\alpha=95\%$, on peut calculer $z_\alpha = 1,96$

- Données : observations x_1, \dots, x_n (et y_1, \dots, y_n)
- Modèle : variables aléatoires X_1, \dots, X_{n1} i.i.d. (et Y_1, \dots, Y_{n2} i.i.d.)
- Erreur possible de décision

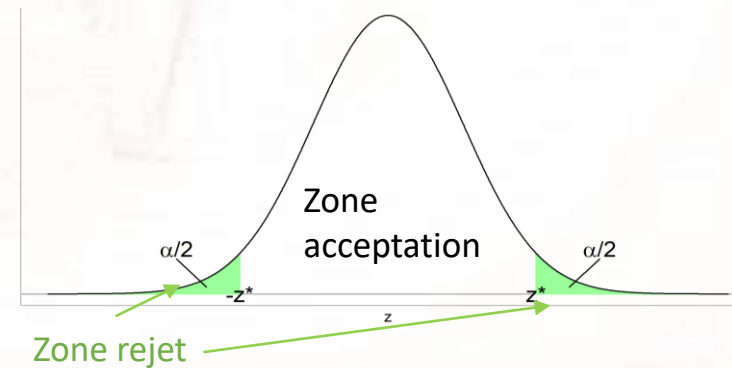
H0 est	Vraie	Fausse
Rejetée	Type I erreur Probabilité α	Décision correcte
Non rejetée	Décision correcte	Type II erreur Probabilité β



- La Puissance du test $= 1 - \beta$ mesure la capacité à empêcher une décision fausse.



La **p value** est la probabilité que la donnée pourrait être observée sous l'hypothèse nulle H_0 . Ici z est la statistique calculée sur l'échantillon.



Ici $z^* = z_{0.95}$ est choisi pour un test avec le risque $\alpha = 0,05$ par exemple.

- Données : observations x_1, \dots, x_n et y_1, \dots, y_n
 - Modèle : variables aléatoires X_1, \dots, X_{n1} i.i.d. et Y_1, \dots, Y_{n2} i.i.d.
-
- Exemples de tests paramétriques
 - Tests pour décider si une moyenne est supérieure à une autre ou une valeur (nommés tests unilatéraux au lieu de bilatéraux considérés ici)
 - Test de normalité pour décider si l'échantillon est gaussien/normal
 - Test de nullité d'un paramètre en régression après estimation du modèle
 - Test d'égalité de variance, ou d'égalité de plus de deux moyennes (anova)
 - Exemple de non paramétrique au lieu de paramétrique
 - Test non paramétrique de comparaison de moyenne
 - Exemple utilisant une loi du chi2
 - Test d'indépendance dans un tableau croisé
 - Test d'adéquation à une loi

- Données : observations x_1, \dots, x_n et y_1, \dots, y_n
 - Modèle : variables aléatoires X_1, \dots, X_{n1} i.i.d. et Y_1, \dots, Y_{n2} i.i.d.
-

Au tableau en cours et en td.