4. Variables aléatoires continues

et leurs distributions

Sam Gyetvay

ECO 2273 – Économetrie I

7 novembre 2025

Variables aléatoires continues

Lors du dernier cours, nous avons étudié les variables aléatoires **discrètes** et leurs distributions

Les variables aléatoires discrètes apparaissent dans *certaines* applications économiques

- Nombre d'entretiens qu'un travailleur recevra s'il envoie n candidatures (binomiale)

Mais les variables de résultat les plus intéressantes en économie sont **continues**

- Revenu, loyer, prix ($)
- Taux d'intérêt, chômage, croissance, inflation, taux de pauvreté (%)
- Productivité, revenu par habitant ($/personne)

Dans le cours d'aujourd'hui, nous définirons la distribution des variables continues à l'aide de leurs **fonctions de densité de probabilité** et de leurs **fonctions de répartition**

Ensuite, nous étudierons quelques distributions continues célèbres

- Uniforme
- Normale
- Chi-carré
- Student-t
- Exponentielle

Fonction de densité de probabilité

Que signifie le fait qu'une variable aléatoire soit continue ?

Une variable aléatoire est continue si elle a une distribution continue

La distribution d'une variable aléatoire continue X est décrite par son **fonction de densité de probabilité** $f(x)$

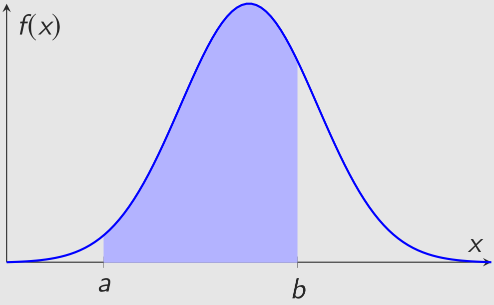
$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

En mots : la probabilité que X soit entre a et b est l'**intégrale** de $f(x)$ entre a et b

Révision : intégration

En calcul, une **intégrale** est l'*aire sous une courbe* :

$$\int_a^b f(x)\, dx = \text{aire sous } f(x) \text{ entre } x = a \text{ et } x = b$$



L'intégration est la version continue de la sommation

Chaque fois que vous voyez un symbole d'intégrale $\int$ (le « S étiré »), ne paniquez pas

Pensez simplement à cela comme la version continue du symbole de sommation $\sum$ (« sigma »)

La définition mathématique d'une intégrale est une sorte de somme spéciale (appelée **somme de Riemann**)

Divisez $[a, b]$ en n sous-intervalles égaux de largeur $\Delta x = \frac{b-a}{n}$

Hauteur du rectangle dans le sous-intervalle i est $f(x_i^*)$, un point quelconque de cet intervalle

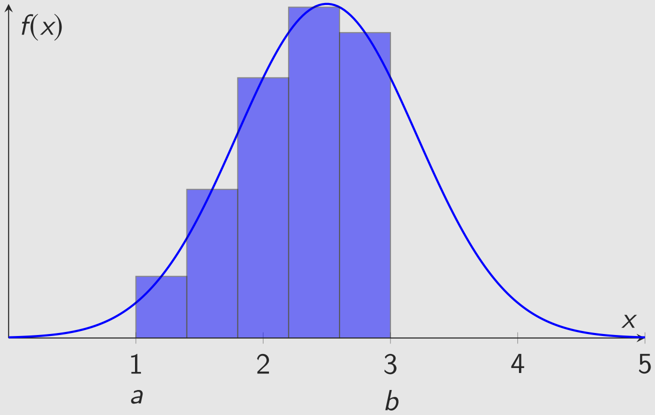
Somme des aires des rectangles :

$$S_n = \sum_{i=1}^{n} f(x_i^*)\Delta x$$

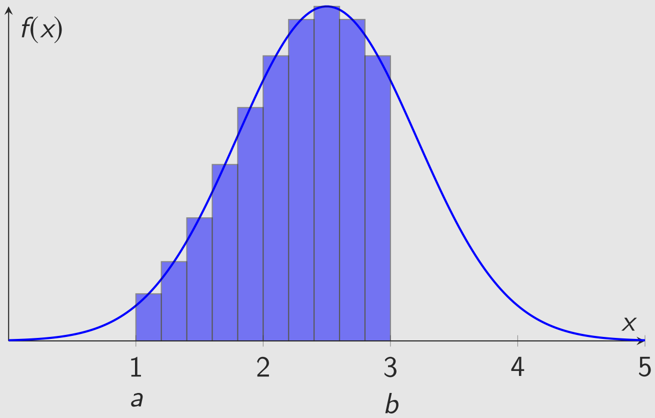
L'intégrale est la « limite » lorsque nous avons un nombre infini de rectangles, chacun étant très petit

$$\int_a^b f(x)\, dx = \lim_{n\to\infty} \sum_{i=1}^{n} f(x_i^*)\Delta x$$

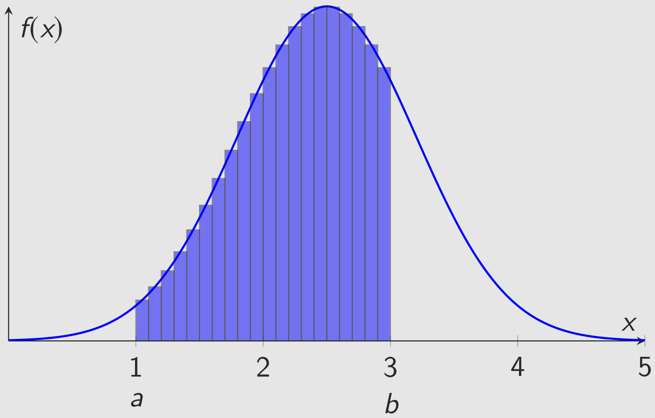
Somme de Riemann ($n = 5$)



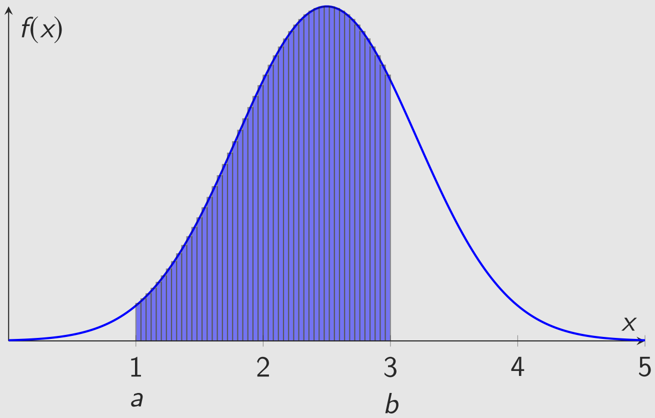
Somme de Riemann ($n = 10$)



Somme de Riemann ($n = 20$)



Somme de Riemann ($n = 50$)



Fonction de masse de probabilité vs fonction de densité de probabilité I

La **fonction de densité de probabilité** $f(x)$ est l'analogue continu de la **fonction de masse de probabilité** $p(x)$ pour les variables aléatoires discrètes

Rappelons les propriétés d'une fonction de masse de probabilité $p(x)$ d'une variable aléatoire discrète X

- $p(x) = P(X = x)$
- $0 \leq p(x) \leq 1$
- $\sum_x p(x) = 1$

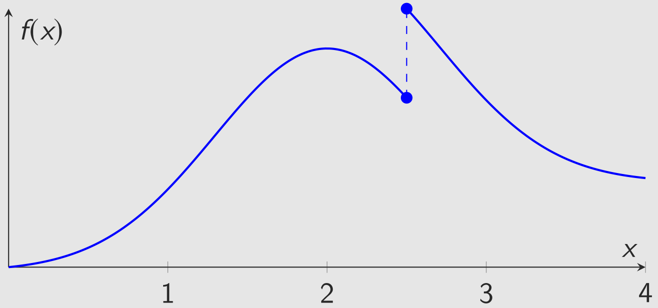
Une fonction de densité de probabilité $f(x)$ a des propriétés similaires définies sur des intervalles $[a, b]$

- $\int_a^b f(x)dx = P(a \leq X \leq b)$
- $f(x) \geq 0$
- $\int_{-\infty}^{+\infty} f(x)dx = 1$

Continuité

La variable aléatoire X est continue si sa fonction de densité de probabilité $f(x)$ est continue

Une fonction $f(x)$ est **continue** en a s'il y a un “saut” soudain ou une “chute” dans $f(x)$ lorsque $x = a$



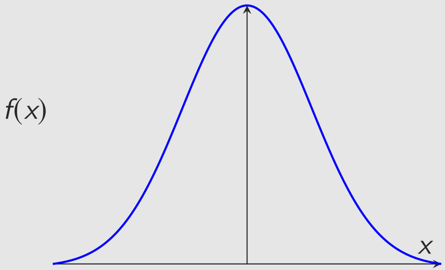
Si $f(x)$ présente un saut en a, alors $P(X = a) > 0$. Par conséquent, une variable aléatoire continue a $P(X = a) = 0$ pour chaque a !

Fonction de masse de probabilité vs fonction de densité de probabilité II

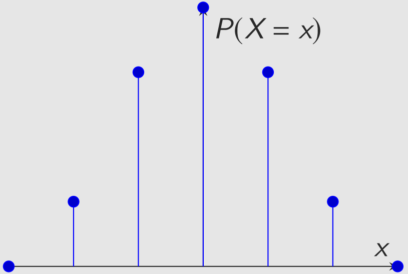
La fonction de densité de probabilité d'une variable aléatoire continue n'a *pas* de points de masse. Sa probabilité n'est non nulle que lorsque nous la calculons sur un intervalle $[a, b]$

La fonction de masse de probabilité d'une variable aléatoire discrète a *uniquement* des points de masse

VA Continue (PDF)



VA Discrète (PMF)



Variables aléatoires mixtes

Les variables aléatoires qui sont continues dans certaines zones mais ont des points de masse sont appelées **variables aléatoires mixtes**

Elles se produisent occasionnellement en économie, mais nous ne les verrons pas dans ce cours, car cela nécessite des mathématiques plus avancées pour les traiter

Un exemple de distribution mixte est le pourcentage de changement dans les revenus d'un individu

$$X = \left(\frac{\text{revenus en année } t - \text{revenus en année } t-1}{\text{revenus en année } t-1} \right) \times 100\%$$

Pour les individus qui étaient employés en $t-1$ mais qui deviennent non employés en t, leurs revenus seront

Variables aléatoires mixtes

Les variables aléatoires qui sont continues dans certaines zones mais ont des points de masse sont appelées **variables aléatoires mixtes**

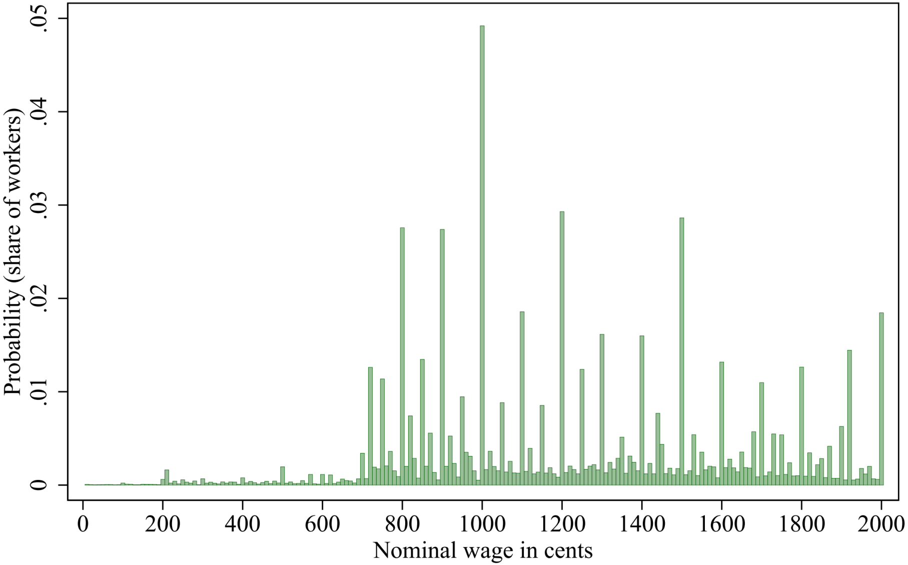
Elles se produisent occasionnellement en économie, mais nous ne les verrons pas dans ce cours, car cela nécessite des mathématiques plus avancées pour les traiter

Un exemple de distribution mixte est le pourcentage de changement dans les revenus d'un individu

$$X = \left(\frac{\text{revenus en année } t - \text{revenus en année } t-1}{\text{revenus en année } t-1} \right) \times 100\%$$

Pour les individus qui étaient employés en $t-1$ mais qui deviennent non employés en t, leurs revenus seront -100%

Un autre exemple est la distribution des salaires horaires. Beaucoup d'employeurs fixent les salaires horaires à des nombres "ronds" (par exemple, 10$ de l'heure, 12$ de l'heure), ce qui crée un "regroupement"



Source : Dube, Naidu et Manning (2025, AER)

Récapitulatif

Une variable aléatoire continue X a une distribution de probabilité continue $f(x)$

Elle n'a pas de points de masse : $P(X = a) = 0$ pour tous les a

Les probabilités ne sont non nulles que pour les intervalles $[a, b]$:

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx \geq 0$$

Les PDF ont des propriétés

- $\int_{-\infty}^{+\infty} f(x)\,dx = 1$
- $f(x) \geq 0$

(Remarquez que nous n'exigeons pas que $f(x) \leq 1$. Pourquoi ?)

Fonction de répartition cumulative

Une autre manière de décrire une distribution est par la **fonction de répartition cumulative** (CDF)

La CDF $F(x)$ mesure la probabilité que X soit inférieur ou égal à x :

$$F(x) = P(X \leq x)$$

La CDF $F(x)$ peut être définie en termes de la PDF $f(x)$ comme

$$F(x) = \int_{-\infty}^{x} f(t)\,dt$$

La CDF a les propriétés

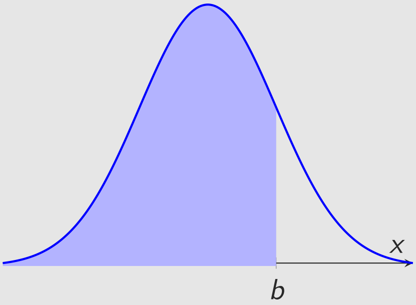
$$0 \leq F(x) \leq 1, \quad \lim_{x \longrightarrow -\infty} F(x) = 0, \quad \lim_{x \longrightarrow +\infty} F(x) = 1$$

Nous pouvons utiliser la CDF pour calculer des probabilités

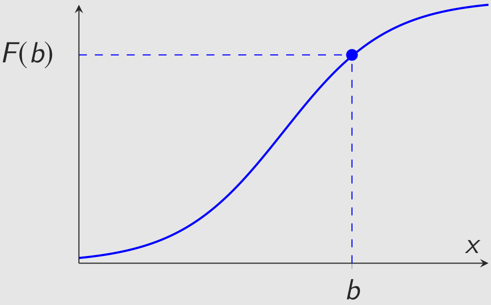
$$P(a \leq X \leq b) = F(b) - F(a)$$

PDF et CDF

PDF de X



CDF de X



Quantiles et la CDF

Rappelez-vous de notre première conférence lorsque nous avons défini les **quantiles**

> *Le quantile p-ième Q_p est la valeur en dessous de laquelle $100 \cdot p\%$ des observations tombent.*

C'est l'inverse de la définition de la CDF !

Si $F(x) = p$, x est le quantile p-ième de X

Si X est continu, F est strictement croissante (pourquoi ?)

Alors $F^{-1}(p) = x$ est unique et égal à Q_p

Tracé de densité

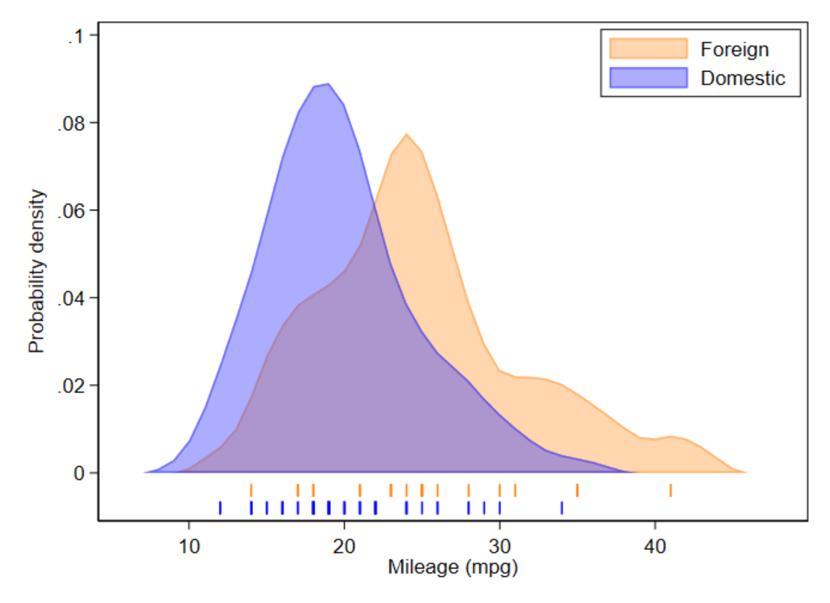
Nous pouvons approximer la PDF d'une variable continue dans un ensemble de données en utilisant une PDF empirique

Les détails de cela dépassent le cadre de ce cours, mais l'idée de base est d'approximer la densité par une combinaison linéaire de fonctions lisses appelées **kernels**

Ces derniers sont connus sous le nom de **tracés de densité de kernel**, et dans STATA, vous pouvez les estimer en utilisant la commande `twoway kdensity`

Les tracés de densité sont comme des histogrammes, mais *lisses*

Tout comme avec les histogrammes où vous devez définir la "largeur de bin", dans les tracés de densité, nous devons définir la "bande passante"



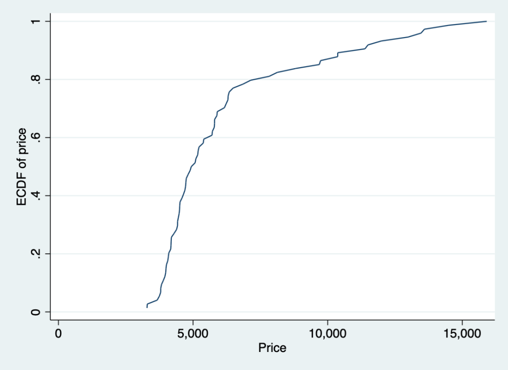
https://www.statalist.org/forums/forum/general-stata-discussion/general/1547538-plot-kernel-density-estimates-as-areas

Fonction de répartition empirique

Tout comme les graphiques de densité, nous pouvons également produire des fonctions de répartition empiriques

Code STATA pour produire une ECDF en utilisant les données `autos.dta` :

```
sysuse auto.dta, clear
sort price
cumul price, gen(ecdf_price)
line ecdf_price price
```



Valeur attendue d'une variable aléatoire continue

Rappelez-vous que la valeur attendue d'une variable aléatoire discrète est

$$E[X] = \sum_x x \cdot p(x)$$

Quelle est la valeur attendue d'une variable aléatoire continue ?

Valeur attendue d'une variable aléatoire continue

Rappelez-vous que la valeur attendue d'une variable aléatoire discrète est

$$E[X] = \sum_{x} x \cdot p(x)$$

Quelle est la valeur attendue d'une variable aléatoire continue ?

(*Indice* : rappelez-vous que l'intégration est la version continue de la sommation)

Valeur attendue d'une variable aléatoire continue

Rappelez-vous que la valeur attendue d'une variable aléatoire discrète est

$$E[X] = \sum_x x \cdot p(x)$$

Quelle est la valeur attendue d'une variable aléatoire continue ?

(*Indice* : rappelez-vous que l'intégration est la version continue de la sommation)

- Remplacez $\sum$ par $\int$
- Remplacez $p(x)$ par $f(x)$
- ... et ajoutez dx

Valeur attendue d'une variable aléatoire continue

Rappelez-vous que la valeur attendue d'une variable aléatoire discrète est

$$E[X] = \sum_x x \cdot p(x)$$

Quelle est la valeur attendue d'une variable aléatoire continue ?

(*Indice* : rappelez-vous que l'intégration est la version continue de la sommation)

- Remplacez $\sum$ par $\int$
- Remplacez $p(x)$ par $f(x)$
- ... et ajoutez dx

$$E[X] = \int_{-\infty}^{+\infty} x f(x)\, dx$$

Variance d'une variable aléatoire continue

Rappelez-vous que la variance d'une variable aléatoire discrète est

$$Var[X] = \sum_x (x - E[X])^2 \cdot p(x)$$

Quelle est la variance d'une variable aléatoire continue ?

Variance d'une variable aléatoire continue

Rappelez-vous que la variance d'une variable aléatoire discrète est

$$Var[X] = \sum_x (x - E[X])^2 \cdot p(x)$$

Quelle est la variance d'une variable aléatoire continue ?

L'intégration est la version continue de la sommation

- Remplacez $\sum$ par $\int$
- Remplacez $p(x)$ par $f(x)$
- ... et ajoutez dx

$$Var[X] = \int_{-\infty}^{+\infty} (x - E[X])^2 f(x) dx$$

Covariance d'une variable aléatoire continue

Rappelez-vous que la covariance de deux variables aléatoires discrètes est

$$Cov(X, Y) = \sum_x \sum_y (x - E[X])(y - E[Y]) \cdot p(x, y)$$

Quelle est la covariance d'une variable aléatoire continue ?

Covariance d'une variable aléatoire continue

Rappelez-vous que la covariance de deux variables aléatoires discrètes est

$$Cov(X, Y) = \sum_x \sum_y (x - E[X])(y - E[Y]) \cdot p(x, y)$$

Quelle est la covariance d'une variable aléatoire continue ?

L'intégration est la version continue de la sommation

- Remplacez $\sum$ par $\int$
- Remplacez $p(x, y)$ par $f(x, y)$
- ... et ajoutez $dxdy$

$$Cov(X, Y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x - E[X])(y - E[Y]) f(x, y) dxdy$$

Exemples de distributions continues

Certaines distributions continues sont célèbres et portent leurs propres noms

Uniforme

- Chaque valeur dans une plage donnée est également probable

Normale (Gaussienne)

- La classique “courbe en cloche” où la plupart des valeurs sont proches de la moyenne

Chi-carré

- Non-négative et asymétrique vers la droite
- Résulte de sommes de normales au carré

Student-t

- Comme une normale mais avec des queues plus épaisses ; utilisée lors de l’estimation de moyennes avec de petits échantillons
- Résulte de la division d’une normale par une Chi-carré

Exponentielle

- Distribution des temps d’attente entre des événements aléatoires
- Similaire à Poisson (distribution d’événements dans une période fixe)

Variables aléatoires uniformes

Une variable aléatoire uniforme $X \sim U(a, b)$ a une densité constante sur l'intervalle $[a, b]$

Fonction de densité de probabilité (PDF) :

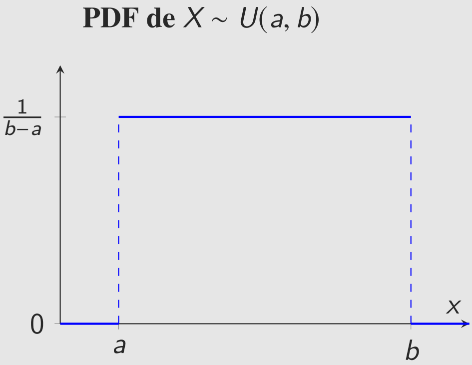
$$f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{sinon} \end{cases}$$

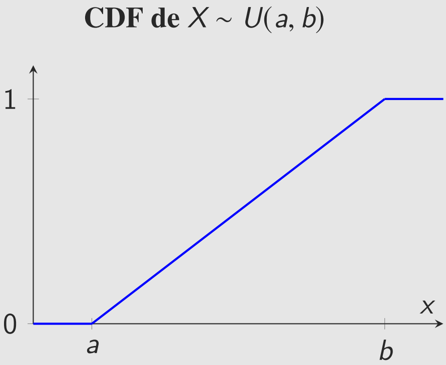
Moyenne : $E[X] = \frac{a+b}{2}$ Variance : $Var(X) = \frac{(b-a)^2}{12}$

Probabilité que X soit dans $[c, d]$, où $a < c < d < b$

$$P(c \leq X \leq d) = \int_c^d f(x)dx = \frac{d-c}{b-a}$$

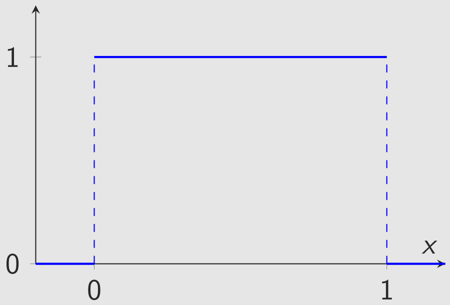
PDF et CDF de $U(a, b)$



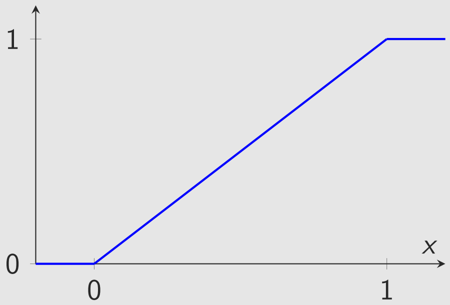


PDF et CDF de $U(0,1)$

PDF de $X \sim U(0,1)$



CDF de $X \sim U(0,1)$



Exemple : Expérience de tarification dynamique d'Uber

Pendant un match de hockey à Montréal, la demande dépasse brièvement l'offre et il y a trop peu de voitures près du centre Bell

Uber souhaite maximiser ses revenus, donc il augmente le prix. En supposant que l'offre de chauffeurs Uber est fixe à court terme, cela a deux effets :

1. les passagers paient plus par trajet,
2. certains passagers décideront que l'Uber est trop cher et prendront le métro à la place

Uber veut connaître le niveau optimal de tarification dynamique, donc il réalise une expérience. Il applique aléatoirement une tarification dynamique entre 5% et 45% uniformément à différents utilisateurs.

Soit $X \sim U(0.05, 0.45)$ le multiplicateur de tarification dynamique

- $E[X] = (0.05 + 0.45)/2 = 0.25$ (surcharge moyenne de 25%)
- $Var(X) = (0.45 - 0.05)^2/12 = 0.16/12 \approx 0.0133$
- $P(0.1 \leq X \leq 0.2) = \frac{0.2-0.1}{0.45-0.05} = 0.25$

Variables aléatoires normales

Une variable aléatoire normale $X \sim N(\mu, \sigma^2)$ a la densité classique en forme de cloche

La plupart des valeurs sont proches de la moyenne μ, et les valeurs éloignées de μ sont de plus en plus improbables

Fonction de densité de probabilité (PDF) :

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

- Symétrique autour de μ (asymétrie = 0)
- Moyenne $E[X] = \mu$, variance $Var(X) = \sigma^2$

Normale standard

Normale standard est le nom spécial donné à une distribution normale de moyenne nulle ($\mu = 0$) et d'écart-type égal à un ($\sigma = 1$)

La normale standard $Z \sim N(0, 1)$ possède également un nom spécial/une lettre grecque pour sa fonction de densité de probabilité $\phi(z)$ et sa fonction de répartition cumulative $\Phi(z)$

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right), \qquad \Phi(z) = \int_{-\infty}^{z} \phi(z)$$

La normale standard est importante car nous pouvons toujours convertir une normale $X \sim N(\mu, \sigma^2)$ en une normale standard en normalisant : soustraire la moyenne et diviser par l'écart-type

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

(Nous utiliserons ce fait lorsque nous construirons des statistiques de test !)

Pourquoi la distribution normale a-t-elle une forme de cloche ?

Rappelez-vous que la densité normale est

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

La partie qui lui donne la “forme de cloche” est $\exp(-x^2)$

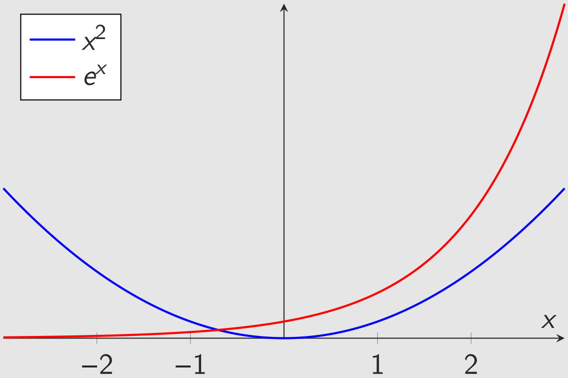
x^2 est un grand nombre lorsque $|x|$ est grand : que x soit un grand nombre positif ou un grand nombre négatif

e^x est très grand pour des valeurs élevées de x, et ≈ 0 pour des valeurs négatives de x

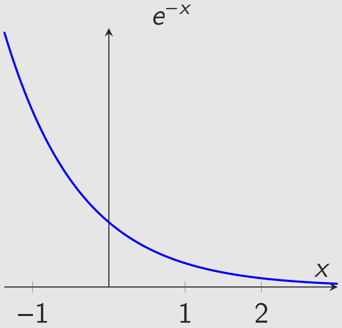
$e^{-x} = \frac{1}{e^x}$ est donc très petit lorsque x est un grand nombre positif, et lorsque x est un grand nombre négatif

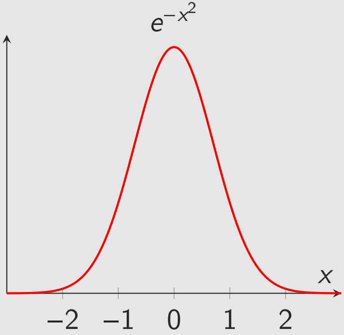
$e^{-x^2} = \frac{1}{e^{x^2}}$ est très petit lorsque x est un grand nombre positif ou un grand nombre négatif, et atteint son maximum à $e^{-x^2} = 1$ lorsque $x = 0$

x^2 et e^x



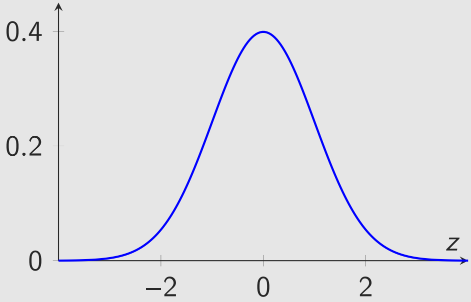
e^{-x} vs e^{-x^2}



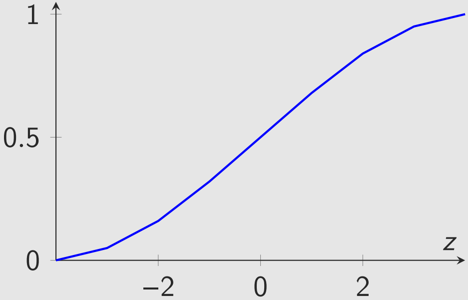


$N(0,1)$ PDF et CDF

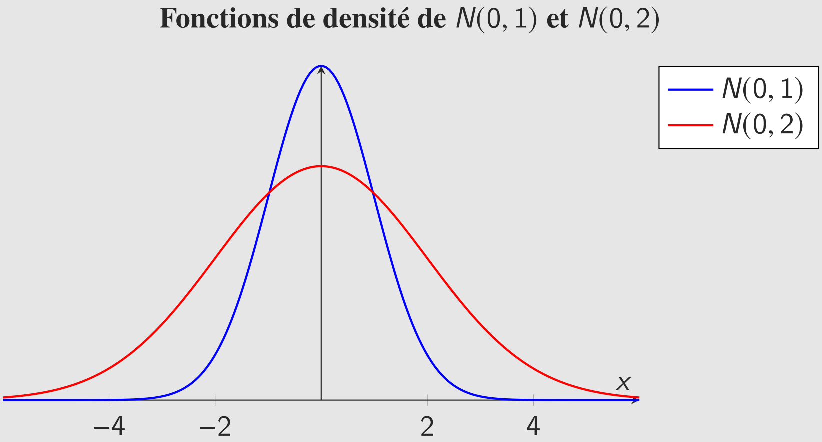
PDF de $Z \sim N(0,1)$



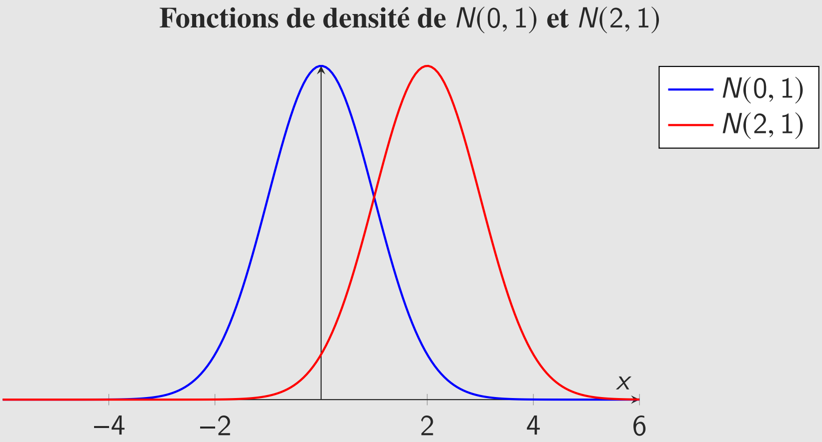
CDF de $Z \sim N(0,1)$



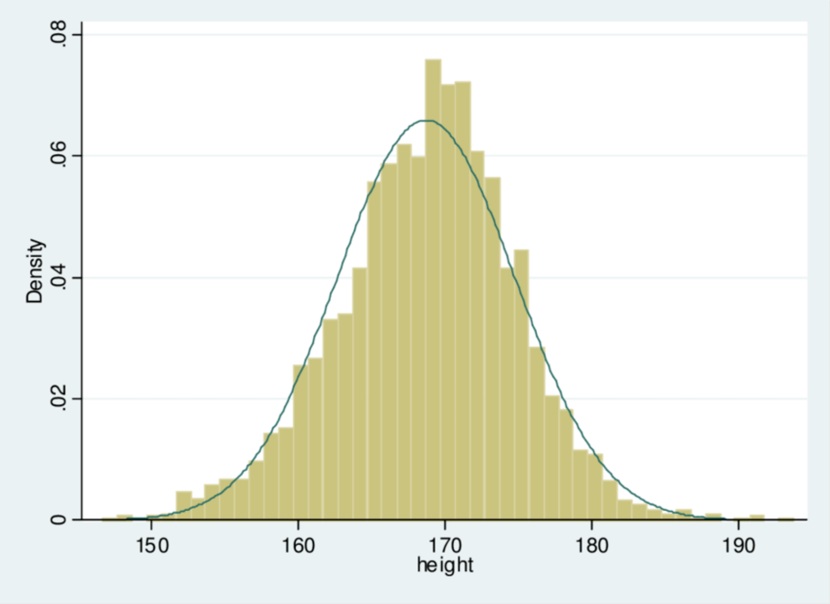
Fonctions de densité normales $N(0,1)$ vs $N(0,2)$



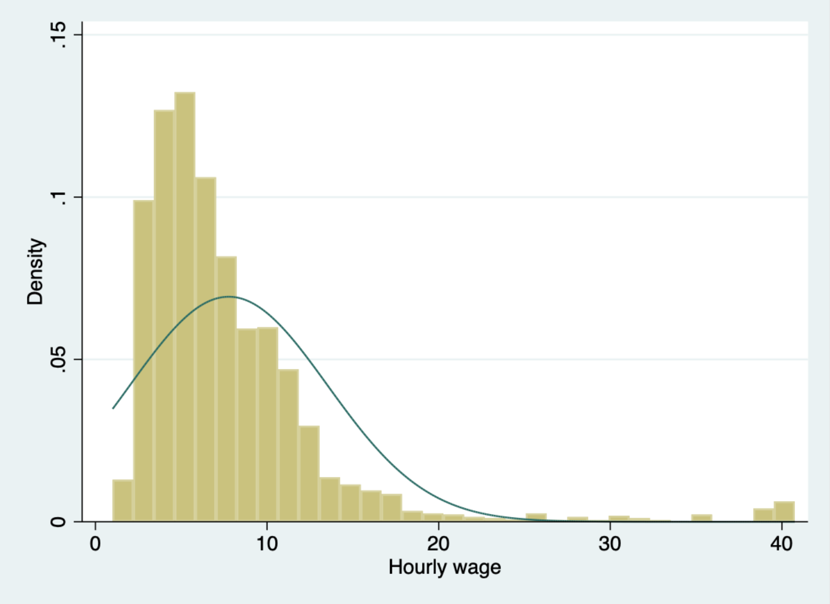
Fonctions de densité normales $N(0,1)$ vs $N(2,1)$



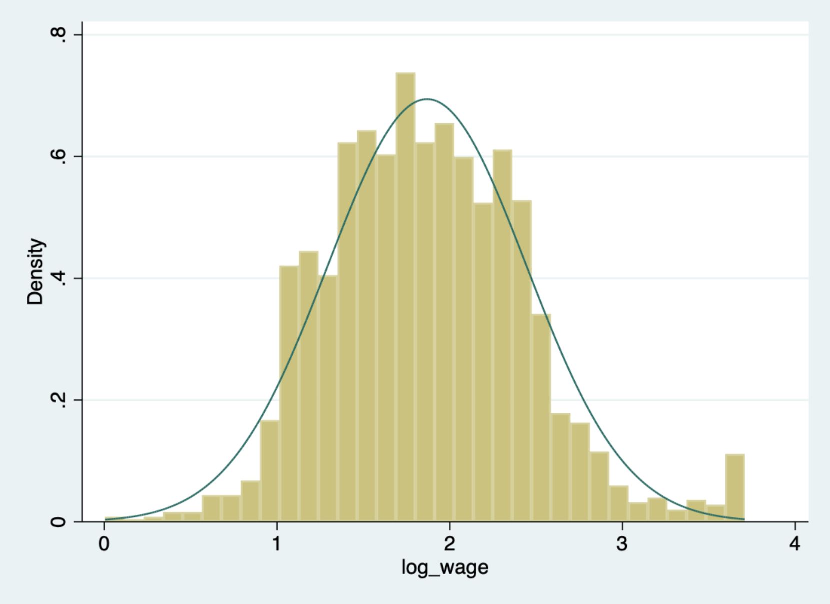
Distribution des tailles des adultes (en cm)



Distribution des salaires horaires dans NLSW88



Distribution des salaires horaires logarithmiques dans NLSW88



Variables aléatoires chi-carré

Une variable aléatoire chi-carré $X \sim \chi^2_k$ provient de la somme des carrés de k normales standard indépendantes :

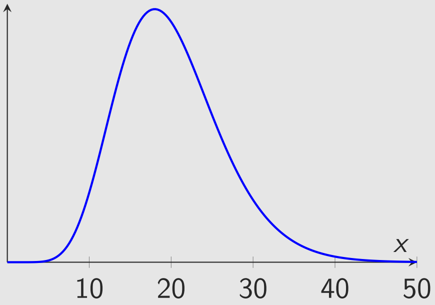
$$X = Z_1^2 + Z_2^2 + \cdots + Z_k^2, \quad Z_i \sim N(0, 1)$$

- Toujours non-négative : $X \geq 0$
- Asymétrique à droite, surtout pour les petits k
- Moyenne $E[X] = k$, variance $Var(X) = 2k$

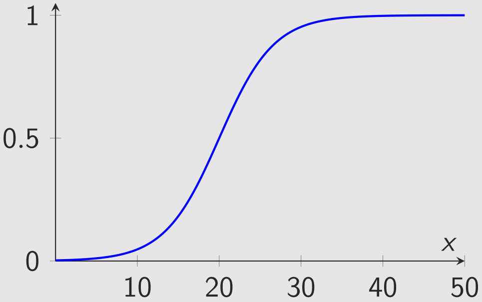
(Nous rencontrerons le chi-carré lorsque nous calculerons des variances)

PDF et CDF de χ^2_{20} (lisse)

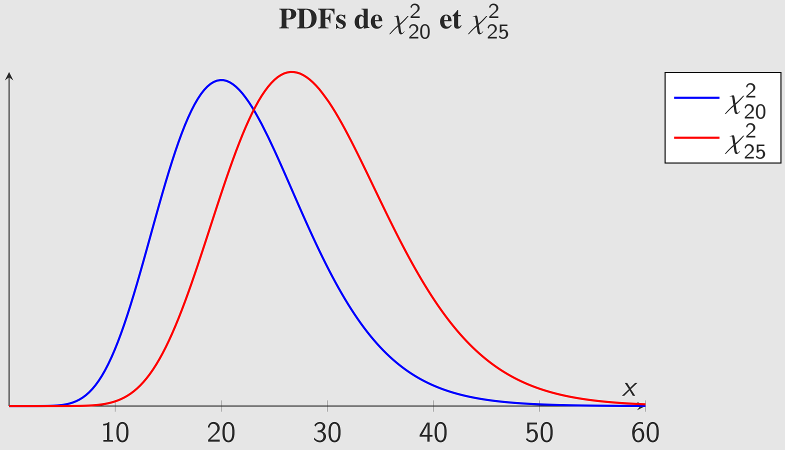
PDF de $X \sim \chi^2_{20}$



CDF de $X \sim \chi^2_{20}$



PDFs de χ^2_{20} vs χ^2_{25} (visuel)



Variables aléatoires Student-t

Une variable aléatoire Student-t X apparaît lorsque vous divisez une variable normale standard $Z \sim N(0,1)$ par une variable χ^2 avec d degrés de liberté

Student-t ressemble à $N(0,1)$, mais elle a des « queues plus épaisses ». Plus d est petit, plus les queues sont épaisses. À mesure que d devient très grand, Student-t se rapproche de la distribution normale

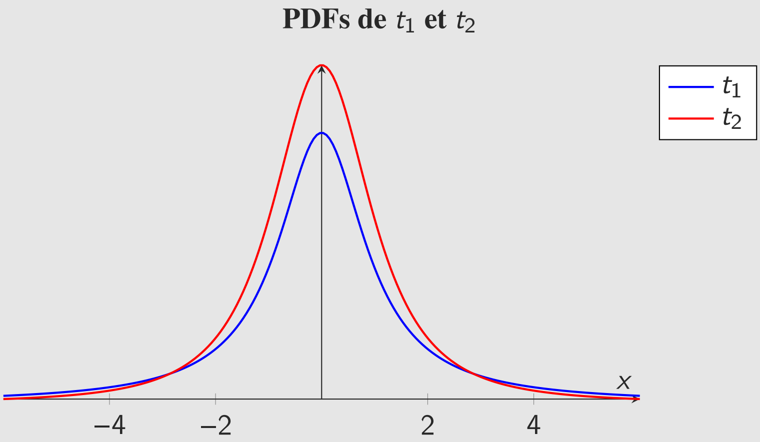
Student-t avec un degré de liberté est appelée une distribution de Cauchy.

Ces distributions dites « à queues épaisses » sont très courantes en finance, où elles sont utilisées pour modéliser des événements extrêmement rares comme un krach boursier.

Dans une distribution « à queues fines » comme la normale, les événements très éloignés (c'est-à-dire à 6 écarts-types ou « six σ » de la moyenne) se produisent très rarement

(Nous rencontrerons Student-t lorsque nous diviserons des normales par leurs variances dans le processus de normalisation)

PDFs de t_1 vs t_2



Variable aléatoire exponentielle

Une variable aléatoire exponentielle $X \sim \mathrm{Exp}(\lambda)$ modélise les temps d'attente entre des événements aléatoires

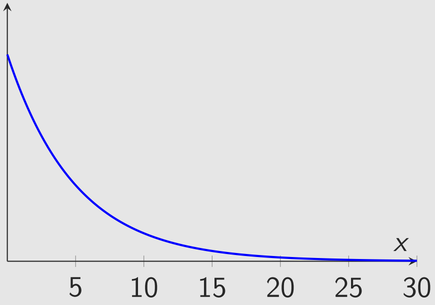
Fonction de densité de probabilité (PDF) :

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0$$

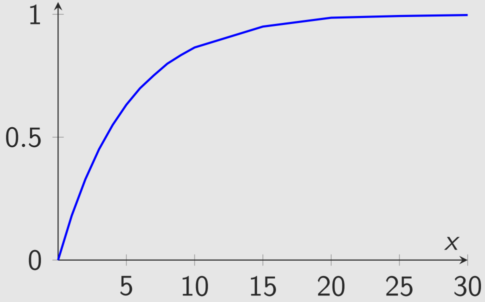
- Moyenne $E[X] = 1/\lambda$, variance $Var(X) = 1/\lambda^2$
- Sans mémoire : $P(X > s + t \mid X > s) = P(X > t)$

$X \sim \mathbf{Exp}(0.2)$ PDF et CDF

PDF de $X \sim \mathbf{Exp}(0.2)$



CDF de $X \sim \mathbf{Exp}(0.2)$



Relation entre les distributions exponentielle et de Poisson

La distribution exponentielle modélise les temps d'attente entre des événements aléatoires, tandis que la distribution de Poisson compte le nombre d'événements dans un intervalle fixe

- Laissons les événements se produire aléatoirement à un taux moyen λ par unité de temps.
- Nombre d'événements dans un intervalle de temps $[0, t]$: $N(t) \sim \text{Poisson}(\lambda t)$
- Temps jusqu'au premier événement : $X \sim \text{Exp}(\lambda)$
- Plus généralement, le temps entre des événements consécutifs est i.i.d. $\text{Exp}(\lambda)$

Poisson compte les événements, l'exponentielle mesure les intervalles de temps entre les événements

Durée du chômage

Supposons que les travailleurs au chômage reçoivent des offres d'emploi qui arrivent indépendamment à un taux moyen constant λ par mois

- Temps d'attente jusqu'à la première offre $\sim \text{Exp}(\lambda)$

Supposons qu'il s'est écoulé un mois et que Sam n'a toujours pas reçu d'offre d'emploi. Le temps d'attente attendu de Sam, à partir d'aujourd'hui, est-il plus court, plus long ou identique à celui qu'il était lorsqu'il a commencé à chercher ?

Durée du chômage

Supposons que les travailleurs au chômage reçoivent des offres d'emploi qui arrivent indépendamment à un taux moyen constant λ par mois

- Temps d'attente jusqu'à la première offre $\sim \text{Exp}(\lambda)$

Supposons qu'il s'est écoulé un mois et que Sam n'a toujours pas reçu d'offre d'emploi. Le temps d'attente attendu de Sam, à partir d'aujourd'hui, est-il plus court, plus long ou identique à celui qu'il était lorsqu'il a commencé à chercher ?

La réponse correcte est : **identique** ! Cela est dû à la propriété sans mémoire de la distribution exponentielle

La probabilité que Sam trouve un emploi dans les prochains mois est la même, que Sam soit au chômage depuis 1 mois ou 12 mois

$$P(X > s + t \mid X > s) = P(X > t)$$

Dépendance à la durée

L'hypothèse de l'absence de mémoire est-elle réaliste ?

Il y a deux façons dont la réalité peut être différente :

1. Le temps jusqu'à la prochaine entrevue augmente avec le temps (les employeurs voient le long intervalle sur le CV de Sam et sentent l'« odeur d'un perdant »)
2. Le temps jusqu'à la prochaine entrevue diminue avec le temps (Sam devient de plus en plus désespéré, postule à des emplois de moins bonne qualité)

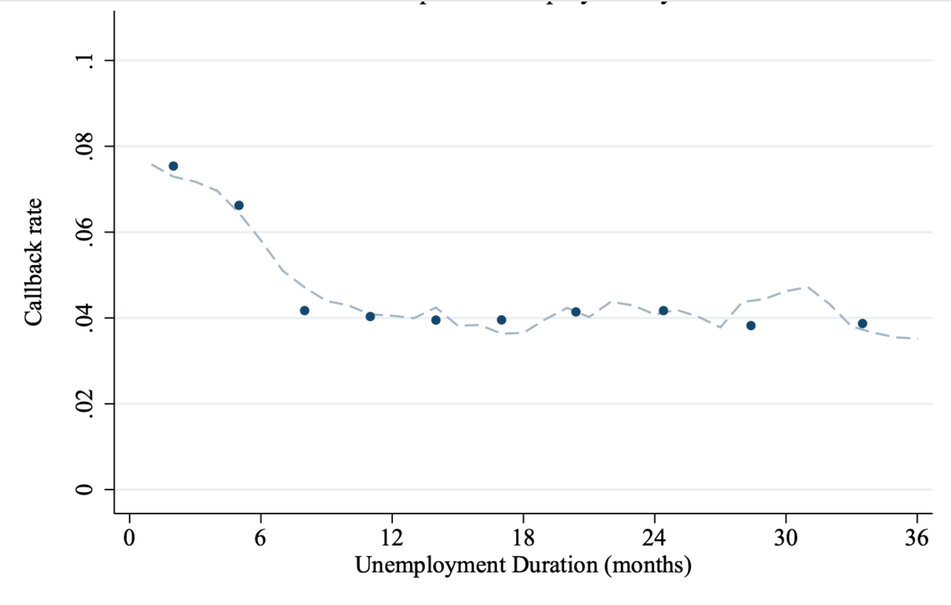
La première est appelée **dépendance à la durée**

Kroft, Lange et Matthew Notowodigdo (QJE, 2013) ont testé la dépendance à la durée en envoyant des milliers de CV fictifs à de vraies offres d'emploi, où ils ont varié aléatoirement l'« intervalle » depuis le dernier emploi sur le CV du travailleur

Ils ont ensuite enregistré quelle fraction de CV a été rappelée pour un entretien

Que pensez-vous qu'ils ont trouvé ?

Kroft, Lange et Notowidigdo (QJE, 2013)



Prochaine classe...

Nous sommes presque à la fin de notre long voyage dans la théorie des probabilités

Dans la prochaine classe, nous étudierons deux sujets qui relient la théorie des probabilités au monde réel et qui constitueront la base de notre étude subséquente de la statistique

Ces sujets sont **l'échantillonnage** et **l'estimation**

1. L'échantillonnage est le processus de prélèvement répété d'observations d'une **population**
2. L'estimation est le processus d'apprentissage sur la distribution dans la population (sa moyenne, sa variance, etc.) à partir de la distribution dans l'échantillon

Pour ce faire, nous utiliserons les deux théorèmes les plus magiques et merveilleux de toute la statistique

- Loi des grands nombres
- Théorème central limite