Révision du niveau 1

Certificat Chargé d'études statistiques

28-29 novembre 2016

Martin Chevalier

Objectifs et organisation

Synthétiser les notions essentielles du niveau 1 du certificat.

Articuler statistique descriptive, statistique inférentielle et applications avec SAS dans une perspective pratique.

Préparer le test d'accès au niveau 2 de la formation.

Objectifs et organisation

Chaque demi-journée est organisée en **trois parties** :

1. Rappels de cours
2. Exercices d'application
3. Mise en œuvre sous SAS

Le fil conducteur de chaque partie est une **étude de cas**, respectivement :

1. Position sur le marché du travail et salaire dans l'**enquête Emploi** de l'Insee ;
2. Performance des élèves aux tests de l'**enquête Pisa** de l'OCDE ;
3. Caractéristiques des prêts à partir de la **base M_CONTRAN** de la Banque de France.

Objectifs et organisation

Analyse univariée : statistiques et représentations

Analyse univariée : inférence

Analyse bivariée : statistiques et représentations

Analyse bivariée : inférence

Analyse univariée : statistiques et représentations

Objectif de la statistique descriptive

L'objectif de la statistique descriptive est de **synthétiser les caractéristiques d'une ou plusieurs variables**.

Ses outils sont particulièrement appropriés pour **communiquer des résultats statistiques** à un **public de non-spécialistes** :

- **représentations graphiques** : diagrammes en bâtons ou circulaires, courbes, nuages de points, etc. ;
- **indicateurs synthétiques** : fréquences, moyennes, indicateurs d'inégalités, etc.

Dans SAS seule une poignée de procédures sont nécessaires pour réaliser tous ces traitements : `PROC GCHART`, `PROC GPLOT`, `PROC FREQ`, `PROC MEANS` principalement.

Étude de cas : l'enquête Emploi en continu

L'ensemble des rappels de cours est structuré autour de l'étude de **deux variables de l'enquête Emploi en continu** :

- la **position sur le marché du travail** au sens du Bureau international du travail : actif occupé, chômeur, inactif ;
- le **salaire mensuel en euros**.

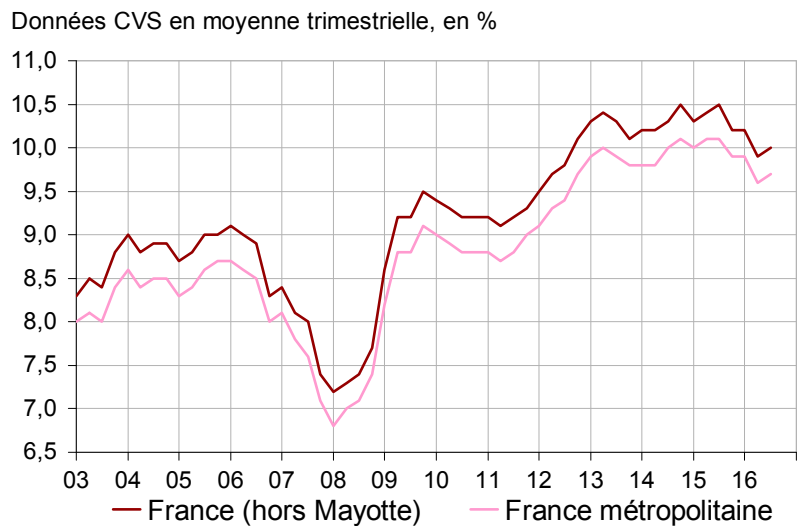
L'enquête Emploi en continu est une enquête permanente de l'Insee dont la principale finalité est de **mesurer le taux de chômage au sens du BIT**.

Tous les traitements présentés ici portent sur 1 752 des quelques 120 000 individus de 15 ans ou plus interrogés au 2012T4.

CEPE

Savoir présenter un tableau ou un grahique

Taux de chômage au sens du BIT



Estimation à +/– 0,3 point près du niveau du taux de chômage et de son évolution d'un trimestre à l'autre
Champ : population des ménages, personnes de 15 ans ou plus
Source : Insee, enquête Emploi

CEPE

De l'importance de la nature des variables

Les outils à mobiliser sont déterminés par la nature des variables à analyser :

- **Variables quantitatives** : leurs modalités peuvent être précisément exprimées les unes en fonction des autres (cardinalité). On distingue les variables discrètes des variables continues.
- **Variables qualitatives** : leurs modalités ne peuvent pas être précisément exprimées les unes en fonction des autres. On distingue les variables ordonnées des variables non-ordonnées.

Exemples Âge (quantitative discrète), salaire (quantitative continue), temps de travail en tranches (qualitative ordonnée), position sur le marché du travail (qualitative non-ordonnée).

Tri à plat : position sur le marché du travail

Le tri à plat permet d'afficher l'**effectif**, la **fréquence** (sous forme de pourcentages) ainsi que l'effectif et la fréquence **cumulés** dans l'ordre des modalités.

Position sur le marché du travail				
ACTEU	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Actif occupé	827	47.20	827	47.20
Chômeur	119	6.79	946	54.00
Inactif	806	46.00	1752	100.00

```
PROC FREQ DATA = d.ee12t4;
    TABLES acteu / MISSING;
    FORMAT acteu $format_acteu.;
RUN;
```

Rappel sur les formats dans SAS

Les formats sont des **tables de correspondance** entre des valeurs et des libellés.

Il est possible de créer un nouvevau format à l'aide de la **PROC FORMAT** :

```
PROC FORMAT;
    VALUE $format_acteu
        "1" = "Actif occupé"
        "2" = "Chômeur"
        "3" = "Inactif"
    ;
RUN;
```

Il est alors possible de l'utiliser dans l'instruction `FORMAT` de la **PROC FREQ** ou de la **PROC GCHART**.

Représentation graphique d'une variable qualitative

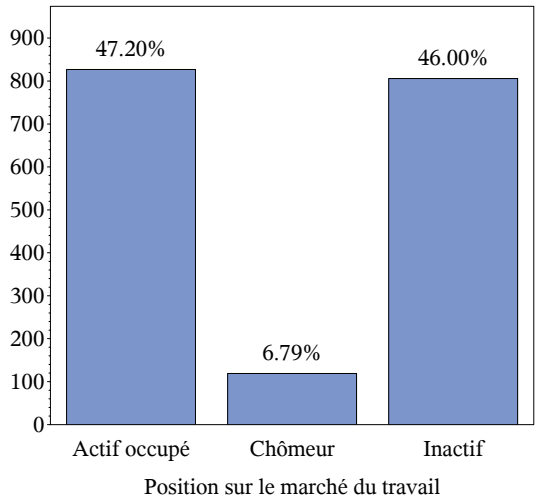
Les deux principales représentations graphiques d'un tri à plat sont les **diagrammes en tuyaux d'orgue et les diagrammes circulaires**.

Les surfaces doivent être **proportionnelles à la fréquence** des modalités qu'elles représentent :

- attention aux représentations qui **ne permettent pas de visualiser correctement les aires** ;
- attention à la **représentation des proportions**.

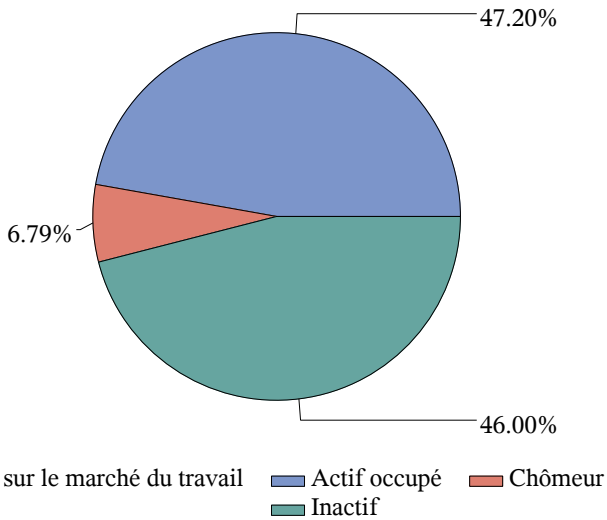
Dans les deux cas, la procédure SAS à utiliser est la **PROC GCHART**.

Diagramme en tuyaux d'orgue : position sur le marché du travail




```
PROC GCHART DATA = d.ee12t4;
    VBAR acteu / PERCENT;
    FORMAT acteu $format_acteu.;
RUN;
```


Diagramme circulaire : position sur le marché du travail




```
PROC GCHART DATA = d.ee12t4;
    PIE acteu / PERCENT = ARROW LEGEND;
    FORMAT acteu $format_acteu.;
RUN;
```

Représentation graphique d'une variable quantitative

L'ensemble de la distribution d'une variable quantitative continue peut être représentée par plusieurs graphiques :

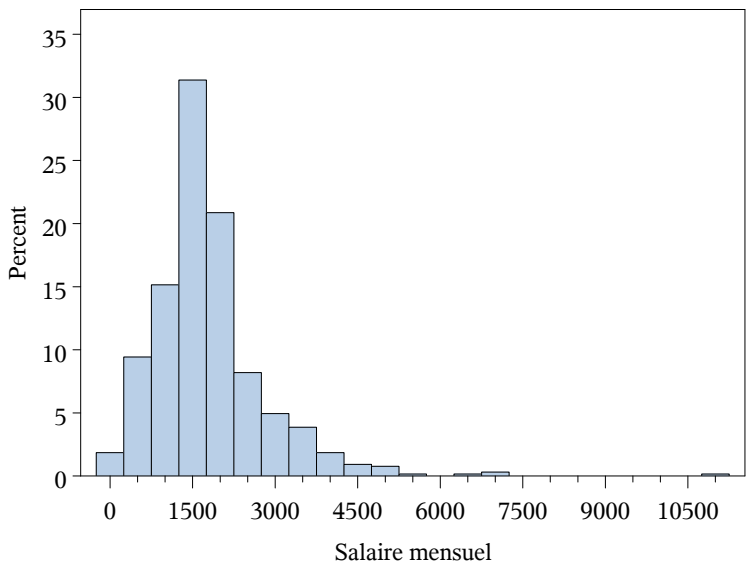
- **Histogramme** : pour un jeu de tranches donné, l'effectif au sein de chaque tranche est représenté par des rectangles.

  Remarque Contrairement au diagramme en bâtons, les tranches ne sont pas pré-déterminées. Le choix de la largeur des tranches modifie l'aspect de l'histogramme.

- **Courbe cumulée croissante** (notée $F(x)$) : pour une valeur donnée de la distribution, elle vaut la part des observations dont la valeur est inférieure.

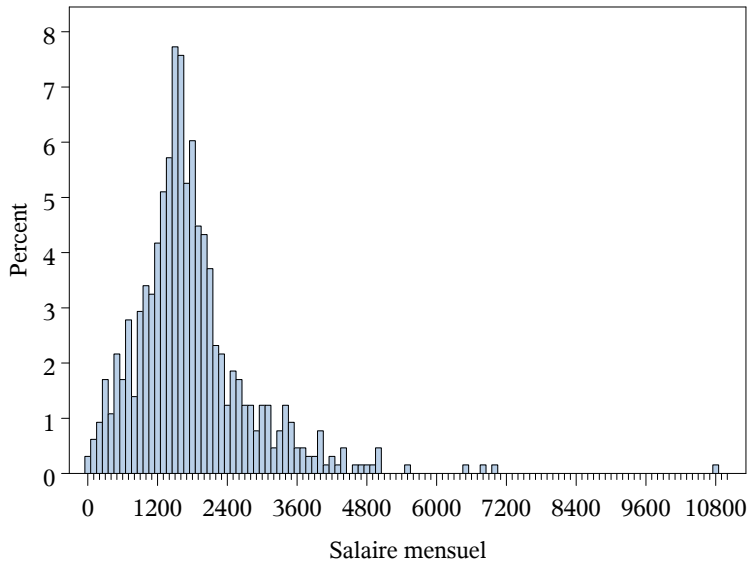
Dans les deux cas, la procédure SAS à utiliser est la **PROC UNIVARIATE**.

Histogramme (1) : salaire



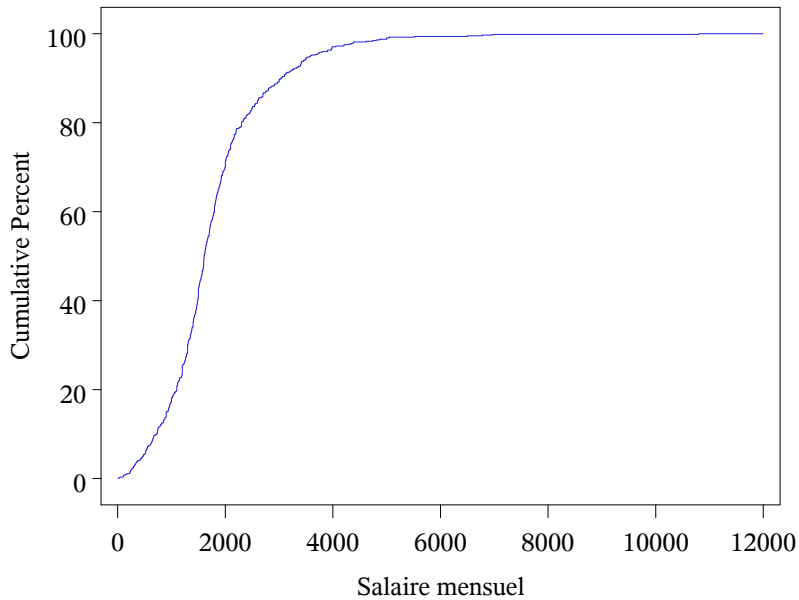

```
PROC UNIVARIATE DATA = d.ee12t4 NOPRINT;
    HISTOGRAM salred / MIDPOINTS = 0 TO 11000
       BY 500;
RUN;
```

Histogramme (2) : salaire




```
PROC UNIVARIATE DATA = d.ee12t4 NOPRINT;
    HISTOGRAM salred / MIDPOINTS = 0 TO 11000
       BY 100;
RUN;
```


Courbe cumulée croissante : salaire




```
PROC UNIVARIATE DATA = d.ee12t4 NOPRINT;
    CDFPLOT salred;
RUN;
```

Caractéristiques de position

Moyenne arithmétique

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

→ Limite : très sensible aux valeurs extrêmes.

Médiane Valeur prise par X qui sépare l'ensemble des individus en deux groupes de même taille.
→ Remarque : se lit directement sur la courbe cumulative.
→ Avantage : très peu sensible aux valeurs extrêmes.

Mode Valeur prise le plus souvent par X.
→ Avantage : très peu sensible aux valeurs extrêmes.
→ Limite : n'est pas toujours défini ni unique.

Caractéristiques de position : salaire

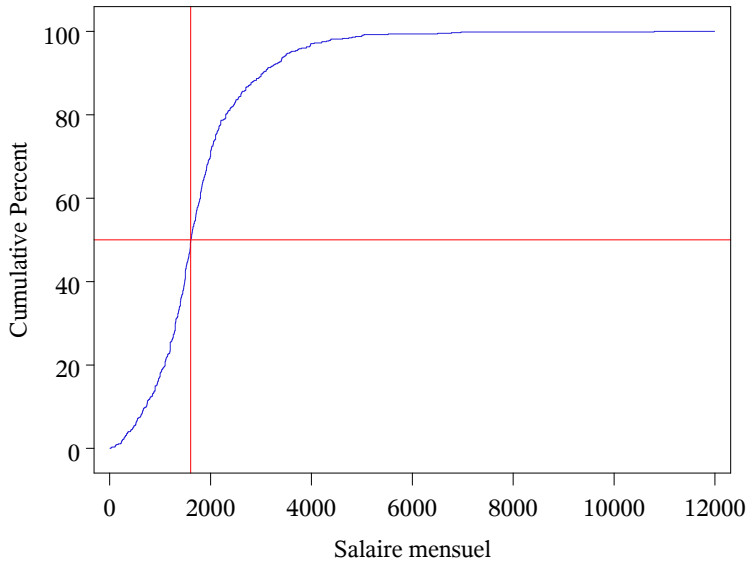
La **PROC MEANS** et la **PROC UNIVARIATE** permettent de calculer les statistiques de position d'une distribution.

Analysis Variable : SALRED Salaire mensuel				
N	N Miss	Mean	Median	Mode
647	1105	1783.92	1607.00	1200.00

```
PROC MEANS DATA = d.ee12t4 N NMISS MEAN MEDIAN
    MODE;
     VAR salred;
RUN;
```

Remarque Ici la distribution est **asymétrique** et comporte quelques valeurs très élevées : moyenne > médiane > mode.

Courbe cumulative croissante et médiane




```
PROC UNIVARIATE DATA = d.ee12t4 NOPRINT;
     CDFPLOT salred / VREF = 50 CV = RED HREF =
        1607 CH = RED;
RUN;
```


Autres types de moyenne

Deux autres types de moyenne peuvent être utilisés dans des situations particulières :

- **Moyenne géométrique** Moyenne de taux de croissance

$$\bar{X}_g = \sum_{i=1}^{k} \sqrt[n_i]{\prod_{i=1}^{k} x_i^{n_i}}$$

  Exemples Augmentation moyenne d'un effectif, taux d'intérêt moyen sur une période.

- **Moyenne harmonique** Moyenne de ratios

$$\bar{X}_h = \frac{\sum_{i=1}^{k} n_i}{\sum_{i=1}^{k} \frac{n_i}{x_i}}$$

  Exemples Vitesse moyenne.

Quantiles d'une distribution

Les quantiles (appelés aussi fractiles) sont construits **à partir de la courbe cumulative croissante**.

Le quantile de niveau α est la plus petite valeur prise par X telle que $F(x) \geq \alpha$.

Exemple Le quantile à 50 % est la plus petite valeur telle que 50 % des valeurs soit inférieures. Il s'agit en réalité de la médiane.

Les **trois quartiles** Q1, Q2 et Q3 séparent la population en quatre sous-ensembles d'effectifs égaux.

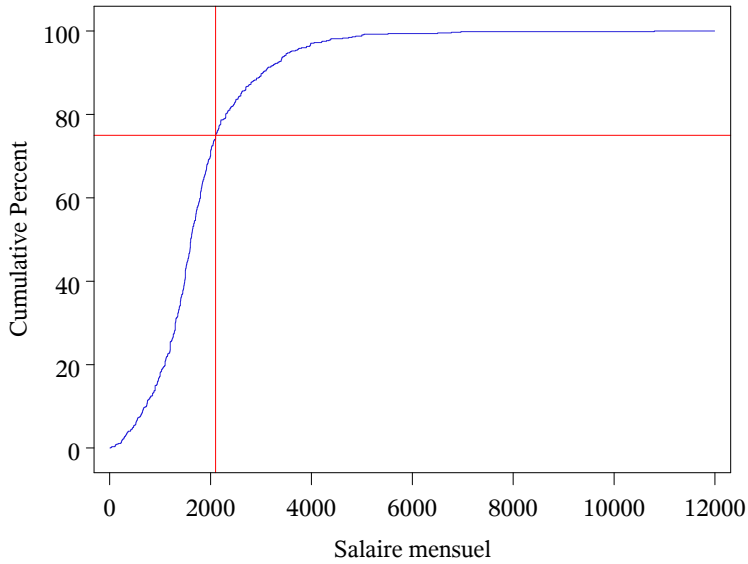
De même, les **neuf déciles** D1, ..., D9 séparent la population en dix sous-ensembles d'effectifs égaux.

Quantiles d'une distribution : salaire

Analysis Variable : SALRED Salaire mensuel						
N	N Miss	10th Pctl	Lower Quartile	Median	Upper Quartile	90th Pctl
647	1105	725.0000000	1200.00	1607.00	2100.00	3037.00

```
PROC MEANS DATA = d.ee12t4 N NMISS P10 Q1
    MEDIAN Q3 P90;
     VAR salred;
RUN;
```

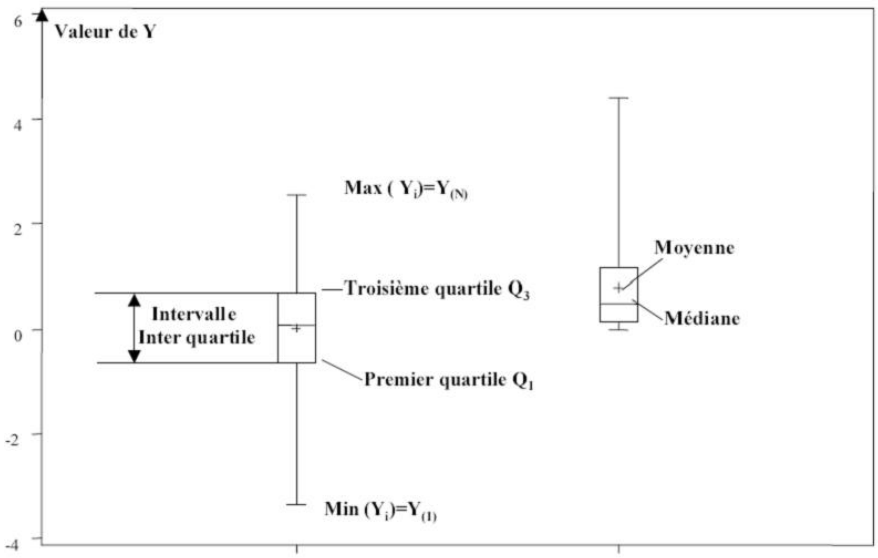
Courbe cumulative croissante et quantiles



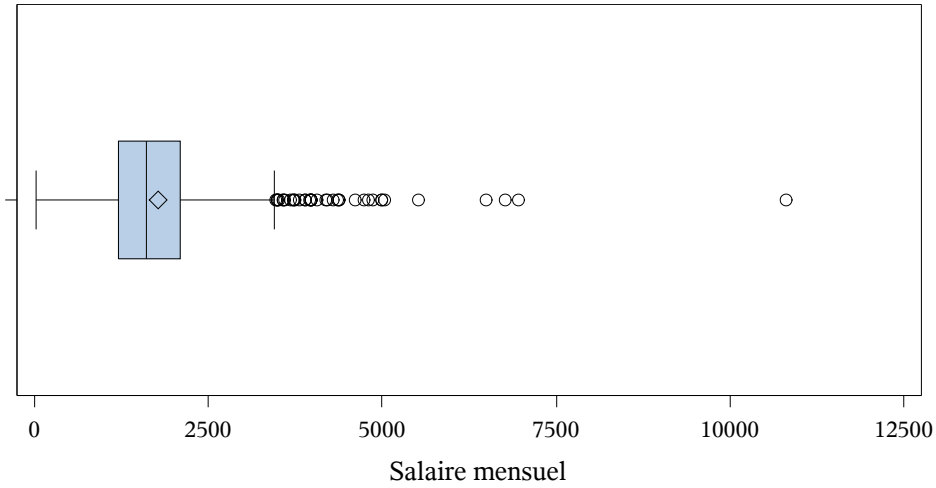

```
PROC UNIVARIATE DATA = d.ee12t4 NOPRINT;
     CDFPLOT salred / VREF = 75 CV = RED HREF50
        = 2100 CH = RED;
RUN;
```


Boîte de Tukey

La boîte de Tukey est une représentation graphique synthétisant plusieurs caractéristiques de position d'une distribution.



Boîte de Tukey : salaire




```
PROC BOXPLOT DATA = d.ee12t4;
    PLOT salred * ensemble / BOXSTYLE =
       SCHEMATIC;
RUN; QUIT;
```


Caractéristiques de dispersion

Les caractéristiques de dispersion rendent compte de façon synthétique de la **variabilité des observations**.

Variance empirique $V(X) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2$

$\rightarrow$ Inconvénient : d'ordre quadratique par rapport à X.

Remarque Le dénominateur de la variance empirique peut être n ou $n-1$ ($n-1$ par défaut dans SAS). En pratique, dès lors que n est grand cela n'affecte que très marginalement la valeur de la variance.

Écart-type $\hat{\sigma}_X = \sqrt{V(X)}$

$\rightarrow$ Avantage : du même ordre de grandeur que X.

Coefficient de variation $CV(X) = \frac{\hat{\sigma}_X}{\bar{X}}$

$\rightarrow$ Avantage : mesure de dispersion relative de X.

Caractéristiques de dispersion : salaire

Analysis Variable : SALRED Salaire mensuel					
N	N Miss	Mean	Variance	Std Dev	Coeff of Variation
647	1105	1783.92	1046558.14	1023.01	57.3462821

```
PROC MEANS DATA = d.ee12t4 N NMISS MEAN VAR
    STD CV;
     VAR salred;
RUN;
```


Indicateurs d'inégalités

On peut construire à partir des quantiles des **indicateurs d'inégalités** :

- Intervalle inter-quartiles : $Q3 - Q1$ ou $\frac{Q3 - Q1}{Max - Min}$;
- Rapport inter-déciles : $D9/D1$
- Autres rapports inter-quantiles : $\frac{Q3}{Q1}$, $\frac{D9}{D5}$, $\frac{D5}{D1}$, etc.

Ces indicateurs rendent compte de l'inégalité dans la distribution de la variable à **différents niveaux**.

La **courbe de Lorenz** et l'**indice de Gini** sont des mesures de concentration qui tiennent compte de l'**ensemble de la distribution**.

Courbe de Lorenz

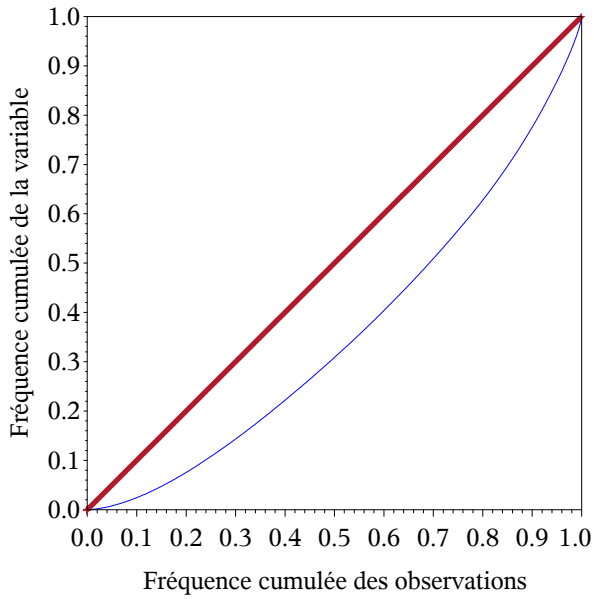
La **courbe de Lorenz** représente la **fréquence cumulée de la variable d'intérêt** en fonction de la **fréquence cumulée des observations**.

Elle permet des interprétations du type : **les 10 % les plus riches possèdent XX % de la richesse totale**.

Interprétation

- Plus la courbe est **proche de la première bissectrice**, moins la répartition est concentrée et plus elle est **égalitaire**.
- Plus la courbe est **proche des côtés du carré**, plus la répartition est concentrée et plus elle est **inégalitaire**.

Courbe de Lorenz : salaire




```
/*Macro-programme spécifique*/
%lorenz(DATA  = d.ee12t4, VAR = salred);
```

Indice de Gini

L'indice de Gini G est un **indice synthétique de concentration** d'une variable additive construit à partir de la courbe de Lorenz.

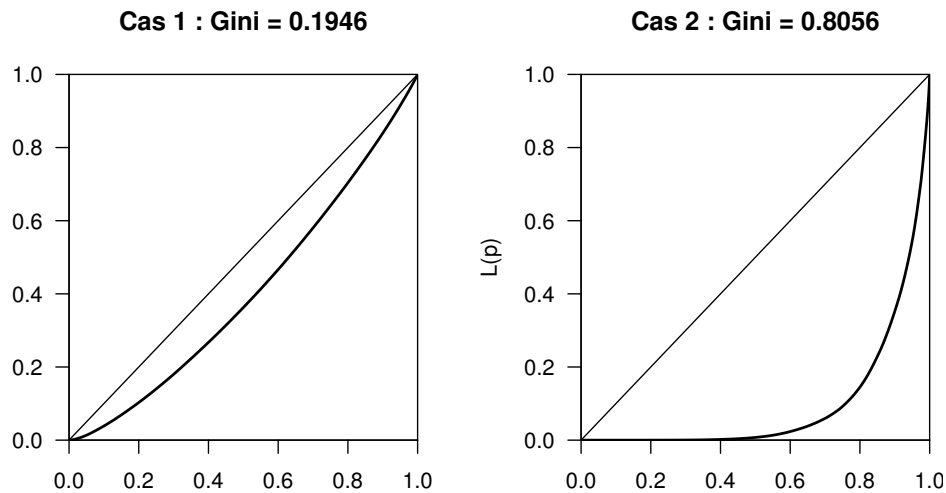
Il correspond à **deux fois l'aire entre la courbe de Lorenz et la première bissectrice** :

- $G = 0$: la courbe de Lorenz est confondue avec la première bissectrice (**égalité parfaite**) ;
- $G = 1$: la courbe de Lorenz est confondue avec les côtés du carré de côté 1 (**inégalité parfaite**).

En pratique, on le calcule à partir de l'aire des trapèzes sous la courbe de Lorenz.

Dans le cas du salaire dans l'enquête Emploi, l'indice de Gini vaut 0,2876.

Courbe de Lorenz et indice de Gini



Remarque Graphiques réalisés à partir de données simulées.

Analyse univariée : inférence

Objectif : De l'échantillon à la population

Les indicateurs standards de la statistique univariée rendent compte de la distribution d'une variable dans les données sur lesquelles ils ont été calculés.

Cependant, bien souvent ces données ne constituent qu'un **échantillon d'une population plus large**, la **population d'inférence**.

L'objectif de la statistique inférentielle est de parvenir, à partir de l'échantillon de données collectées, à une **estimation dans la population d'inférence** :

- construction d'**intervalles de confiance** autour des valeurs estimées dans l'échantillon ;
- mise en œuvre de **tests statistiques** faisant intervenir une ou plusieurs variables (*cf.* partie 4).

Démarche : Postuler un modèle

Pour répondre à ces objectifs, la statistique inférentielle formule un certain nombre d'**hypothèses** sur les données.

Chaque observation de la variable à analyser est vue comme une **réalisation indépendante** d'une **suite de variables aléatoires de même loi**.

Une variable aléatoire est une **fonction qui associe à chaque événement** issu d'une expérience aléatoire une **valeur numérique**.

Si la loi suivie par les variables aléatoires est connue (*i.e.* tabulée), alors il est possible de **déterminer le comportement des statistiques calculées à partir des données**.

Exemple : Jet d'un dé parfaitement équilibré

L'ensemble des éventualités est : $\Omega = \{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\}$ et la probabilité de chacune est $1/6$. Dans ce contexte, on peut assez naturellement définir la variable aléatoire X :

$$X : \begin{array}{ccc} \Omega & \to & \{1,2,3,4,5,6\} \\ \omega & \mapsto & \begin{cases} 1 \text{ si } \boxdot, 2 \text{ si } \boxdot, 3 \text{ si } \boxdot, \\ 4 \text{ si } \boxdot, 5 \text{ si } \boxdot, 6 \text{ si } \boxdot \end{cases} \end{array}$$

On sait alors (par définition) que la variable aléatoire X suit une loi uniforme sur $\{1,2,3,4,5,6\}$.

Remarque On aurait aussi bien pu définir la variable aléatoire $\tilde{X}$:

$$\tilde{X} : \begin{array}{ccc} \Omega & \to & \{0,1\} \\ \omega & \mapsto & \begin{cases} 0 \text{ si } \boxdot, \boxdot \text{ ou } \boxdot \\ 1 \text{ si } \boxdot, \boxdot \text{ ou } \boxdot \end{cases} \end{array}$$

Caractéristiques d'une variable aléatoire

Fonction de répartition : en notant $\mathbb{P}$ la loi de probabilité associée à la variable aléatoire X, on note F sa fonction de répartition définie par :

$$\begin{array}{rccc} & \mathbb{R} & \to & [0;1] \\ F & : & a & \mapsto & \mathbb{P}(X \leq a) \end{array}$$

Remarques La **fonction cumulative croissante** est la contrepartie empirique de la fonction de répartition. La **fonction quantile** est la réciproque de la fonction de répartition.

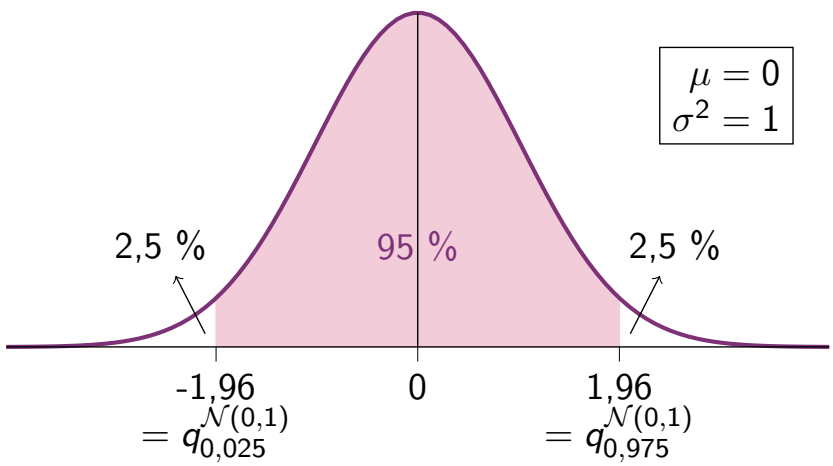
L'**espérance** et la **variance** sont définies par (cas d'une variable aléatoire discrète prenant K valeurs distinctes) :

$$E(X) = \sum_{k=1}^{K} k \times \mathbb{P}(X = k) \quad \text{et} \quad V(X) = \sum_{k=1}^{K} (k - E(X))^2 \times \mathbb{P}(X = k)$$

Quelques lois à connaître (1) : Loi normale

Loi normale (ou gaussienne) La loi normale d'espérance μ et de variance σ^2 est notée $\mathcal{N}(\mu, \sigma^2)$. Sa fonction de répartition est :

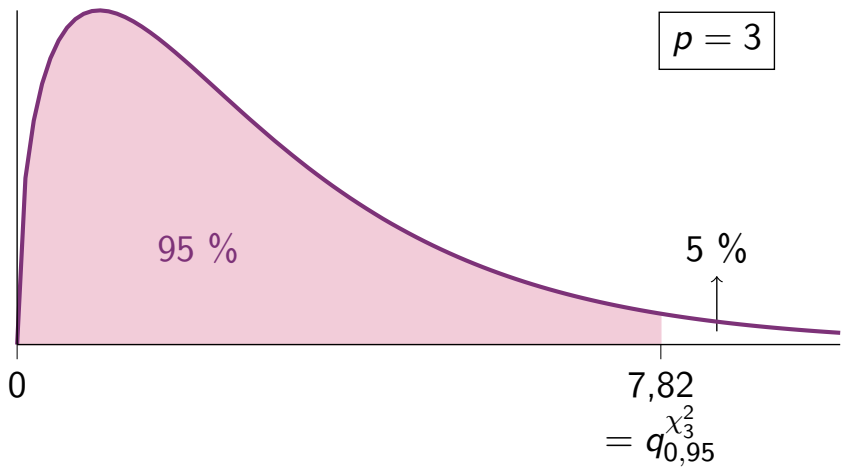
$$F(a) = \int_{-\infty}^{a} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$



Quelques lois à connaître (2) : Loi du χ^2

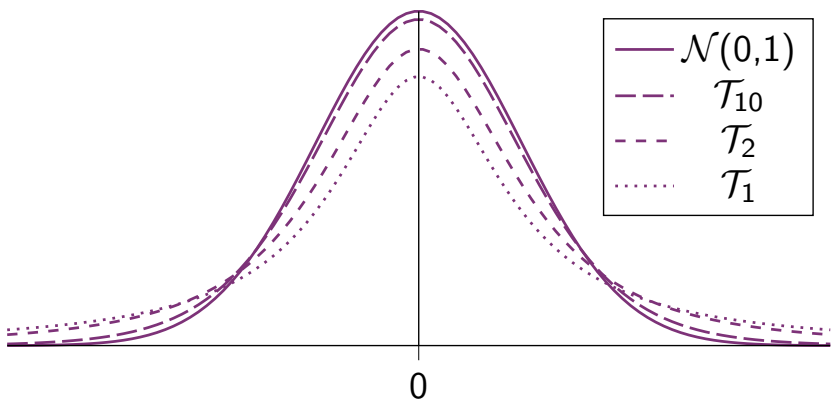
La loi du χ^2 à p degrés de liberté est notée χ^2_p. Si $X_1, \ldots, X_p$ sont indépendantes et suivent une loi normale centrée réduite $\mathcal{N}(0,1)$, alors

$$\sum_{k=1}^{p} X_k^2 \hookrightarrow \chi^2_p$$



Quelques lois à connaître (3) : Loi de Student

La loi de Student à p degrés de liberté est notée $\mathcal{T}_p$. Si $Y \hookrightarrow \mathcal{N}(0,1)$ et $X \hookrightarrow \chi^2_p$ sont indépendantes, alors $Z = \frac{Y}{\sqrt{X/p}} \hookrightarrow \mathcal{T}_p$.



Remarque Si $Z \hookrightarrow \mathcal{T}_p$ alors $Z \xrightarrow[p\to+\infty]{\mathcal{L}} \mathcal{N}(0,1)$

Théorème central limite (1)

Soit $P \times n$ variables aléatoires

$$X_1^{(1)}, \ldots, X_n^{(1)}, \quad X_1^{(2)}, \ldots, X_n^{(2)}, \quad \ldots, \quad X_1^{(P)}, \ldots, X_n^{(P)}$$

indépendantes et suivant une même loi quelconque d'espérance μ et de variance σ^2 finies avec $\sigma^2 \neq 0$.

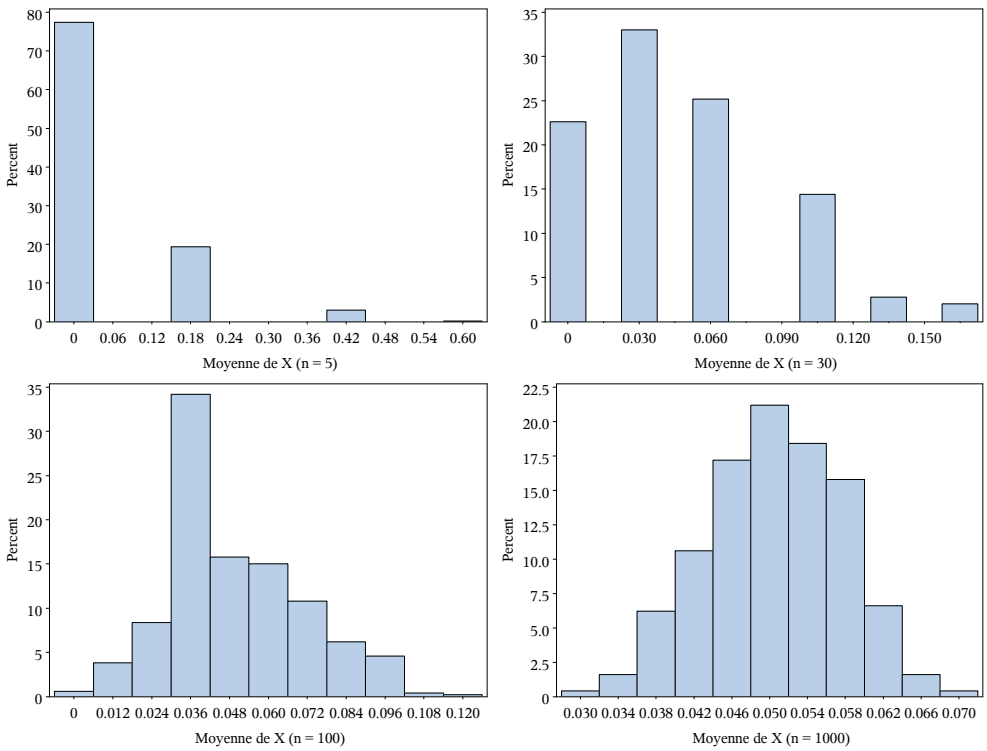
Pour chacun des P groupes de taille n, on calcule la moyenne

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

Si on note $\bar{X}$ la variable aléatoire correspondant à la distribution des P moyennes ainsi calculées, le théorème central limite énonce que :

$$\bar{X} \xrightarrow[n \to +\infty]{\mathcal{L}} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Théorème central limite (2)



Estimation par intervalle d'une moyenne (1)

En raison des fluctuations d'échantillonnage, la **moyenne empirique** calculée dans l'échantillon peut être **relativement éloignée de la moyenne dans la population**.

Il est néanmoins possible de construire un **intervalle de confiance** de la moyenne :

- plus l'intervalle est large, plus on a de chances que la vraie valeur soit effectivement à l'intérieur...
- ... mais plus l'intervalle est large, moins il est informatif.

En pratique, **les seuils retenus sont 90 %, 95 % et 99 %**.

On retient ainsi que pour $n > 30$ l'**intervalle de confiance à 95 % de la moyenne** est :

$$IC_{95\ \%}(\bar{X}) = \left[\bar{X} - 1,96 \times \frac{\hat{\sigma}_X}{\sqrt{n}}; \bar{X} + 1,96 \times \frac{\hat{\sigma}_X}{\sqrt{n}}\right]$$

$se_X = \hat{\sigma}_X/\sqrt{n}$ est appelée l'**erreur standard** .

Estimation par intervalle d'une moyenne (2)

Ce résultat est fondé sur le **théorème central limite**.

Quelle que soit la distribution de X et sous les hypothèses du théorème central limite, on sait que :

$$\bar{X} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

En **centrant** et en **réduisant** cette variable aléatoire, on obtient :

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1)$$

Estimation par intervalle d'une moyenne (3)

En pratique, **l'écart-type théorique σ est inconnu** : il est estimée par l'**écart-type empirique** $\hat{\sigma}_X$:

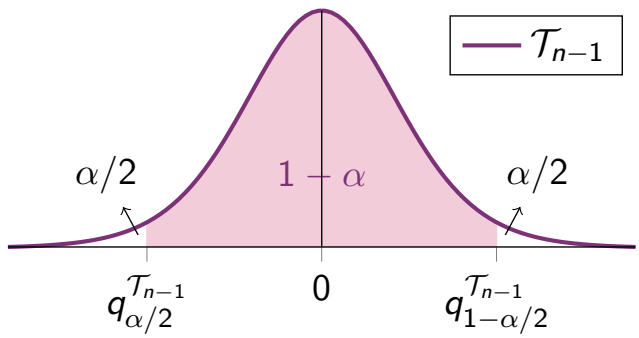
$$\hat{\sigma}_X = \sqrt{V(X)} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2}$$

On peut alors montrer que :

$$\frac{\bar{X} - \mu}{\hat{\sigma}_X/\sqrt{n}} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{T}_{n-1}$$

C'est dont à partir des **quantiles de la loi de Student** à $n-1$ degrés de liberté qu'est construit l'intervalle de confiance de la moyenne.

Estimation par intervalle d'une moyenne (4)



$$\begin{aligned}
1-\alpha &= \mathbb{P}\left(q^{\mathcal{T}_{n-1}}_{\alpha/2} \leq \frac{\bar{X}-\mu}{\hat{\sigma}_X/\sqrt{n}} \leq q^{\mathcal{T}_{n-1}}_{1-\alpha/2}\right) \\
&= \mathbb{P}\left(-q^{\mathcal{T}_{n-1}}_{1-\alpha/2} \leq \frac{\bar{X}-\mu}{\hat{\sigma}_X/\sqrt{n}} \leq q^{\mathcal{T}_{n-1}}_{1-\alpha/2}\right) \\
&= \mathbb{P}\left(\bar{X} - q^{\mathcal{T}_{n-1}}_{1-\alpha/2}\frac{\hat{\sigma}_X}{\sqrt{n}} \leq \mu \leq \bar{X} + q^{\mathcal{T}_{n-1}}_{1-\alpha/2}\frac{\hat{\sigma}_X}{\sqrt{n}}\right)
\end{aligned}$$

d'où : $IC_{1-\alpha\ \%}(\bar{X}) = \left[\bar{X} - q^{\mathcal{T}_{n-1}}_{1-\alpha/2}\frac{\hat{\sigma}_X}{\sqrt{n}}; \bar{X} + q^{\mathcal{T}_{n-1}}_{1-\alpha/2} \times \frac{\hat{\sigma}_X}{\sqrt{n}}\right]$

Estimation par intervalle du salaire moyen (1)

Sur les 647 observations de l'EEC 2012T4 utilisées, la variable de salaire a une moyenne $\bar{X}$ de 1 784 € et un écart-type empirique $\hat{\sigma}_X$ de 1 023 €.

On cherche à construire un intervalle de confiance à 95 % : $\alpha = 0,05$ et le quantile à $1 - \alpha/2 = 0,975$ d'une loi de Student à $647 - 1 = 646$ degrés de liberté vaut environ $1,96$.

$$IC_{95\ \%}(\bar{X}) = \left[1\ 784 - 1,96 \times \frac{1\ 023}{\sqrt{647}}; 1\ 784 + 1,96 \times \frac{1\ 023}{\sqrt{647}}\right]$$

$$IC_{95\ \%}(\bar{X}) = [1\ 705\ €; 1\ 863\ €]$$

Estimation par intervalle du salaire moyen (2)

La **PROC MEANS** permet de calculer automatiquement un intervalle de confiance de la moyenne au seuil désiré avec les mots-clés STDERR, CLM et ALPHA :

Analysis Variable : SALRED Salaire mensuel						
N	N Miss	Mean	Std Dev	Std Error	Lower 95% CL for Mean	Upper 95% CL for Mean
647	1105	1783.92	1023.01	40.2188399	1704.95	1862.90

```
PROC MEANS DATA = d.ee12t4 N NMISS MEAN STD
    STDERR CLM ALPHA = 0.05;
     VAR salred;
RUN;
```

Analyse bivariée : statistiques et représentations

Objectifs de la statistique bivariée

La statistique univariée synthétise les caractéristiques d'une variable prise **isolément**.

Cependant, dans la plupart des cas c'est la **relation entre plusieurs variables** qui constitue le principal enjeu de l'étude.

Exemples Diplôme et position sur le marché du travail, âge et salaire.

Les outils de la statistique bivariée permettent de **traiter ce type de questions**.

Le niveau 2 de la formation est en grande partie consacré aux méthodes d'analyse multivariées (plus de deux variables) :

- Module 2 : **analyse de données** (4-5 janvier 2017) ;
- Modules 3 et 4 : **régression** (30 janvier-1er février et 1-3 mars 2017).

Croisement de deux variables qualitatives

Le **tri croisé** (ou **tableau de contingence**) correspond à la ventilation des observations selon les modalités de deux variables qualitatives X et Y. Il comporte :

- les **effectifs de cellule** (n_{ij}) : nombre d'individus présentant la modalité i de X et la modalité j de Y ;
- les **pourcentages de cellule** (p_{ij}) : part des individus présentant la modalité i de X et la modalité j de Y ;
- les effectifs et pourcentages **marginaux** ($n_{i\cdot}$, $p_{i\cdot}$, $n_{\cdot j}$, $p_{\cdot j}$) : distribution univariée de X et de Y respectivement ;
- les pourcentages **en ligne** (ou **profil-ligne**, $p_{j|i}$) : distribution de Y pour la modalité i de X ;
- les pourcentages **en colonne** (ou **profil-colonne**, $p_{i|j}$) : distribution de X pour la modalité j de Y.

Ces statistiques sont calculées par la **PROC FREQ**.

Tri croisé : diplôme et position sur le marché du travail

Frequency

Table of ACTEU by sup

ACTEU(Position sur le marché du travail)	sup(Diplômé du supérieur)		
	0	1	Total
Actif occupé	545	282	827
Chômeur	94	25	119
Inactif	706	100	806
Total	1345	407	1752

```
PROC FREQ DATA = d.ee12t4;
    TABLES acteu*sup / NOCOL NOROW NOPERCENT;
    FORMAT acteu $format_acteu.;
RUN;
```

Tri croisé : diplôme et position sur le marché du travail

Percent

Table of ACTEU by sup			
ACTEU(Position sur le marché du travail)	sup(Diplômé du supérieur)		
	0	1	Total
Actif occupé	31.11	16.10	47.20
Chômeur	5.37	1.43	6.79
Inactif	40.30	5.71	46.00
Total	1345 76.77	407 23.23	1752 100.00

```
PROC FREQ DATA = d.ee12t4;
    TABLES acteu*sup / NOCOL NOROW NOFREQ;
    FORMAT acteu $format_acteu.;
RUN;
```


Tri croisé : diplôme et position sur le marché du travail

Percent
Row Pct

Table of ACTEU by sup			
ACTEU(Position sur le marché du travail)	sup(Diplômé du supérieur)		
	0	1	Total
Actif occupé	31.11 65.90	16.10 34.10	47.20
Chômeur	5.37 78.99	1.43 21.01	6.79
Inactif	40.30 87.59	5.71 12.41	46.00
Total	1345 76.77	407 23.23	1752 100.00

Tri croisé : diplôme et position sur le marché du travail

Percent Col Pct	Table of ACTEU by sup		
ACTEU(Position sur le marché du travail)	sup(Diplômé du supérieur)		
	0	1	Total
Actif occupé	31.11 40.52	16.10 69.29	47.20
Chômeur	5.37 6.99	1.43 6.14	6.79
Inactif	40.30 52.49	5.71 24.57	46.00
Total	1345 76.77	407 23.23	1752 100.00

Mise en évidence de sur- ou sous-représentations

La comparaison des pourcentages en ligne ou en colonne avec les pourcentages marginaux permet de mettre en évidence des **sur- ou sous-représentations**.

Exemples Surreprésentation des personnes diplômés du supérieur parmi les actifs occupés et sous-représentation parmi les inactifs.

Cette analyse permet de mesurer le **lien entre les variables** : des sur- ou sous-représentations manifestes indiquent des **écarts à la situation d'indépendance**.

Situation d'indépendance

On peut montrer qu'en cas d'**indépendance totale des deux variables**, l'effectif de chaque cellule devrait être le **produit des effectifs marginaux rapporté au nombre total d'observations** :

$$n^*_{ij} = \frac{n_{i.} \times n_{.j}}{n}$$

Pour chaque cellule, il est donc possible de comparer les effectifs observés n_{ij} et les effectifs théoriques sous l'hypothèse d'indépendance n^*_{ij}.

Intuition : **Plus les effectifs observés diffèrent des effectifs théoriques sous l'hypothèse d'indépendance**, plus il est probable que les variables soient **statistiquement liées**.

Effectifs observés et effectifs théoriques : diplôme et position sur le marché du travail

Frequency Expected

Table of ACTEU by sup

ACTEU(Position sur le marché du travail)	sup(Diplômé du supérieur)		
	0	1	Total
Actif occupé	545 634.88	282 192.12	827
Chômeur	94 91.356	25 27.644	119
Inactif	706 618.76	100 187.24	806
Total	1345	407	1752

```
PROC FREQ DATA = d.ee12t4;
    TABLES acteu*sup / EXPECTED NOCOL NOROW
       NOPERCENT;
    FORMAT acteu $format_acteu.;
RUN;
```

Statistique du χ^2

La statistique (ou distance) du χ^2 fait ainsi intervenir la **somme des carrés des écarts entre effectifs théoriques et effectifs observés** :

$$D^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - n_{ij}^*)^2}{n_{ij}^*}$$

où les variables X et Y ont respectivement I et J modalités.

Propriétés

- $D^2 = 0$: les deux variables sont indépendantes ;
- **Plus D^2 est grand, plus on est fondé à penser que les deux variables sont liées**.

Remarque C'est la **loi** que suit D^2 sous l'hypothèse d'indépendance qui permet de déterminer le « seuil » à partir duquel on peut affirmer que les variables X et Y sont statistiquement liées (*cf.* partie 4).

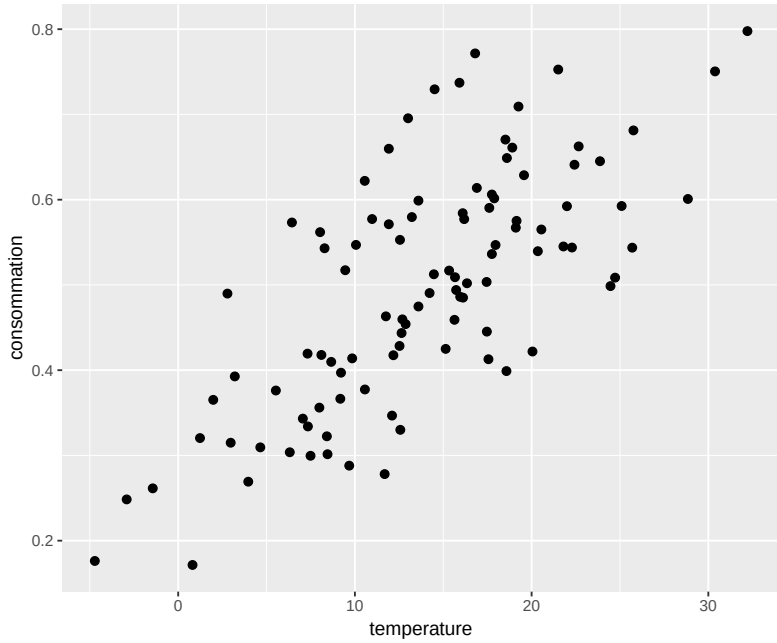
Contribution au χ^2 : diplôme et position sur le marché du travail

Cell Chi-Square	Table of ACTEU by sup		
ACTEU(Position sur le marché du travail)	sup(Diplômé du supérieur)		
	0	1	Total
Actif occupé	12.725	42.052	
Chômeur	0.0765	0.253	
Inactif	12.3	40.646	
Total	1345	407	1752

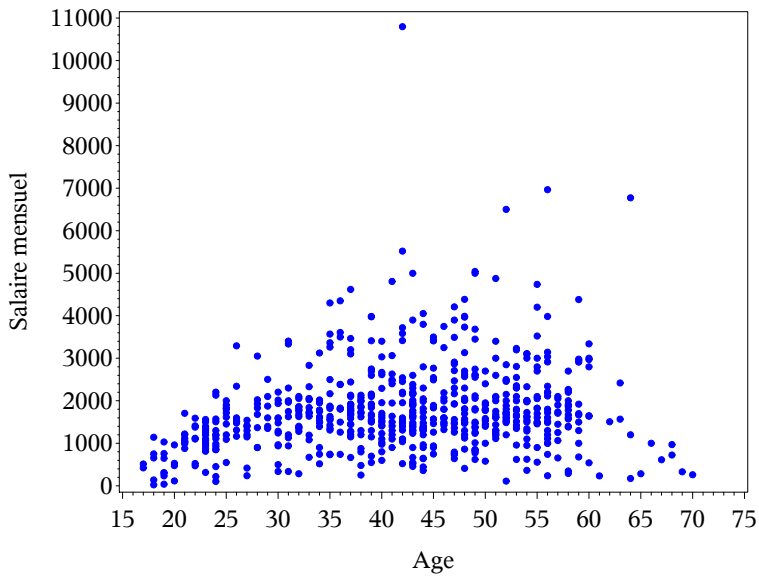
```
PROC FREQ DATA = d.ee12t4;
    TABLES acteu*sup / CELLCHI2 NOFREQ NOCOL
       NOROW NOPERCENT;
    FORMAT acteu $format_acteu.;
RUN;
```

Nuage de points de deux variables quantitatives

La relation entre deux variables quantitatives peut être représentée par un **nuage de points**.



Nuage de points : âge et salaire mensuel




```
PROC GPLOT DATA = d.ee12t4;
    PLOT salred*age / HAXIS = AXIS1 VAXIS =
       AXIS2;
RUN; QUIT;
```

Rappel sur les paramètres graphiques

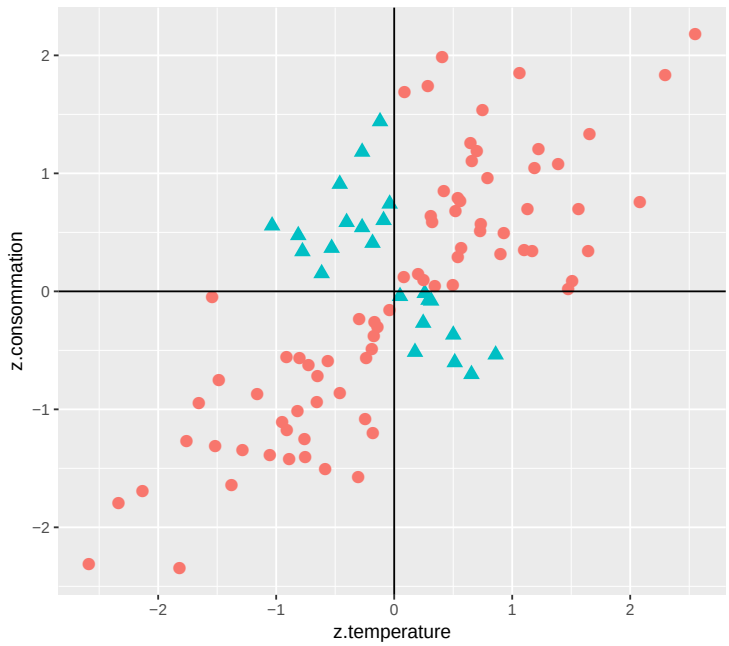
Dans SAS les mots-clés `AXIS` et `SYMBOL` permettent de paramétrer les axes et les marqueurs utilisés dans les graphiques respectivement.

Les paramètres utilisés pour produire le graphique de la diapositive précédente sont :

```
AXIS1 ORDER = (15 TO 75 BY 5);
AXIS2 LABEL=(ANGLE = 90) ;
SYMBOL1 C = BLUE VALUE = DOT I = NONE;
```


Nuage de points des variables centrées réduites

Centrer et réduire les variables facilite la visualisation des **variations concomittantes**.



Covariance de deux variables quantitatives

Définition La covariance empirique de deux variables X et Y est définie par

$$Cov(X, Y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{X})(y_i - \bar{y})$$

Remarque Sur la diapositive précédente, les points représentés par des disques contribuent positivement à la covariance, ceux représentés par des triangles négativement.

Interprétation

- $Cov(X, Y) > 0$: X et Y varient dans le même sens ;
- $Cov(X, Y) < 0$: X et Y varient dans des sens contraires.

Limite La valeur de la covariance n'est pas bornée.

Coefficient de corrélation de Pearson

Définition Le coefficient de corrélation linéaire de Bravais-Pearson $r_{X,Y}$ des variables X et Y est défini par :

$$r_{X,Y} = \frac{Cov(X, Y)}{\sqrt{V(X)V(Y)}} = \frac{Cov(X, Y)}{\hat{\sigma}_X \hat{\sigma}_Y}$$

Remarque Le coefficient de corrélation linéaire est une normalisation de la covariance.

Interprétation

- $r_{X,Y}$ proche de 1 : X et Y varient dans le même sens ;
- $r_{X,Y}$ proche de -1 : X et Y varient dans des sens contraires.

Limite $r_{X,Y}$ est très sensible aux valeurs extrêmes.

Coefficient de corrélation des rangs de Spearman

La sensibilité aux valeurs extrêmes du coefficient de corrélation de Pearson conduit à lui chercher des **alternatives**.

Le **coefficient de corrélation des rangs** de Spearman en est une :

$$r^S_{X,Y} = r_{R_X,R_Y} = \frac{Cov(R_X, R_Y)}{\sqrt{V(R_X)V(R_Y)}}$$

où R_X et R_Y sont les rangs des observations pour les variables X et Y respectivement.

Comme $r_{X,Y}$, $-1 \leq r^S_{X,Y} \leq 1$ mais $r^S_{X,Y}$ est **beaucoup moins sensible aux valeurs extrêmes**.

Corrélation entre âge et salaire

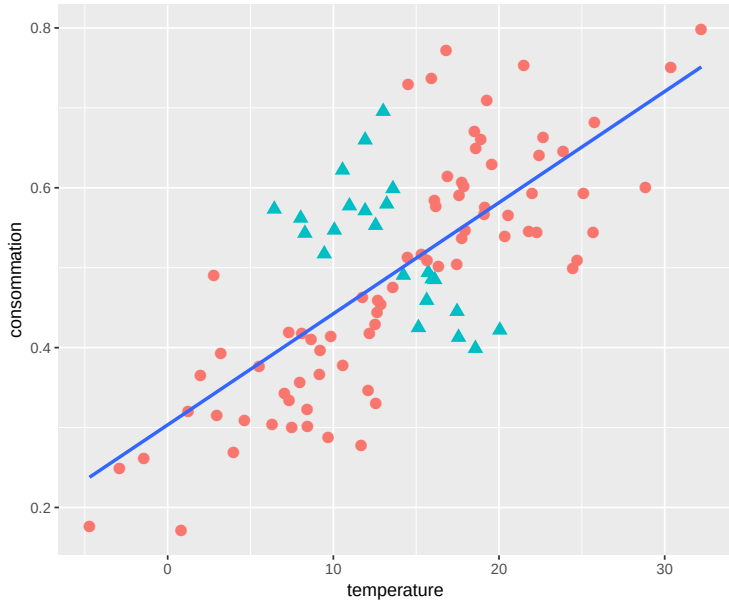
Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations		
	age	**SALRED**
age Age	1.00000 1752	0.21024 <.0001 647
SALRED Salaire mensuel	0.21024 <.0001 647	1.00000 647

Spearman Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations		
	age	**SALRED**
age Age	1.00000 1752	0.21383 <.0001 647
SALRED Salaire mensuel	0.21383 <.0001 647	1.00000 647

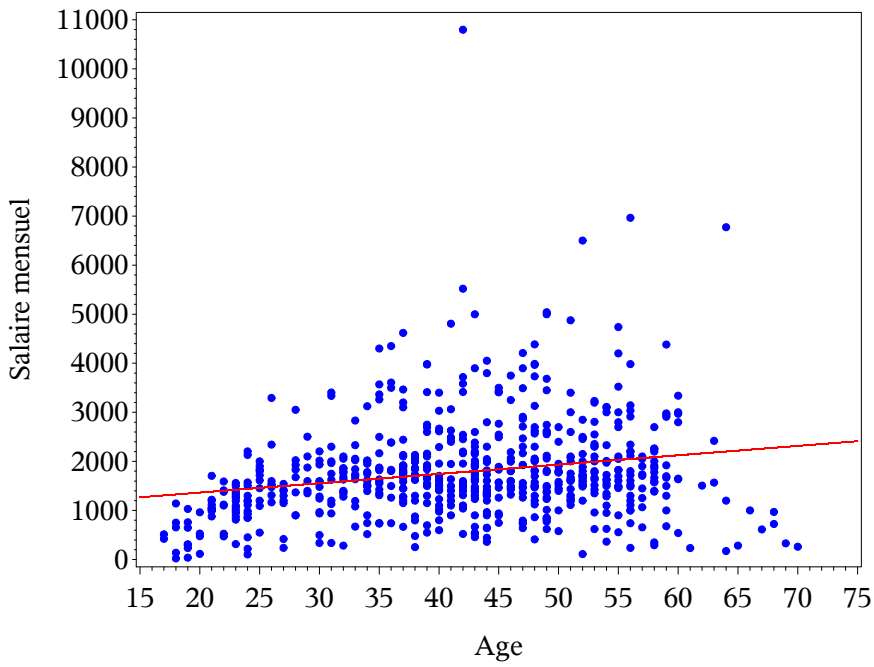
```
PROC CORR DATA = d.ee12t4 PEARSON SPEARMAN;
    VAR age salred;
RUN;
```

Droite de régression

Il est également possible de faire apparaître sur le nuage de points la **droite de la régression linéaire** correspondante (*cf.* module 3, 30 janvier-1er février 2017).



Droite de régression : âge et salaire (1)



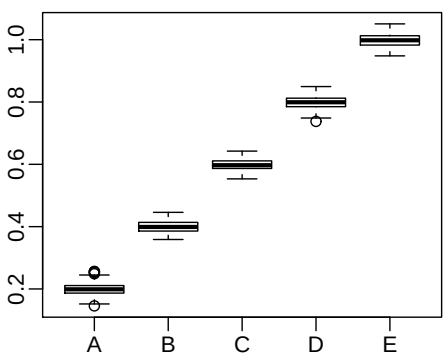
Droite de régression : âge et salaire (2)

```
/*Régression linéaire*/
PROC REG DATA = d.ee12t4 NOPRINT;
    MODEL salred = age;
    OUTPUT OUT = d.ee12t4 P = salredhat;
RUN; QUIT;

/*Options graphiques*/
AXIS1 ORDER = (15 TO 75 BY 5);
AXIS2 LABEL=(ANGLE = 90) ;
SYMBOL1 C = BLUE VALUE = DOT I = NONE;
SYMBOL2 C = RED VALUE = NONE I = JOIN;

/*Nuage de points*/
PROC GPLOT DATA = d.ee12t4;
    PLOT salred*age=1 salredhat*age=2 / HAXIS =
        AXIS1 VAXIS = AXIS2 OVERLAY;
RUN; QUIT;
```

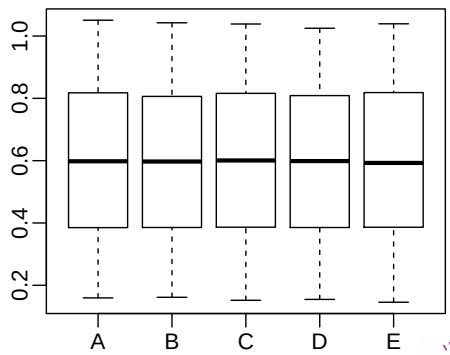

Représenter la relation entre une variable qualitative et une variable quantitative

Pour représenter la relation entre une variable qualitative X et une variable quantitative Y, il est possible de représenter plusieurs boîtes de Tukey côte-à-côte.

Liaison parfaite

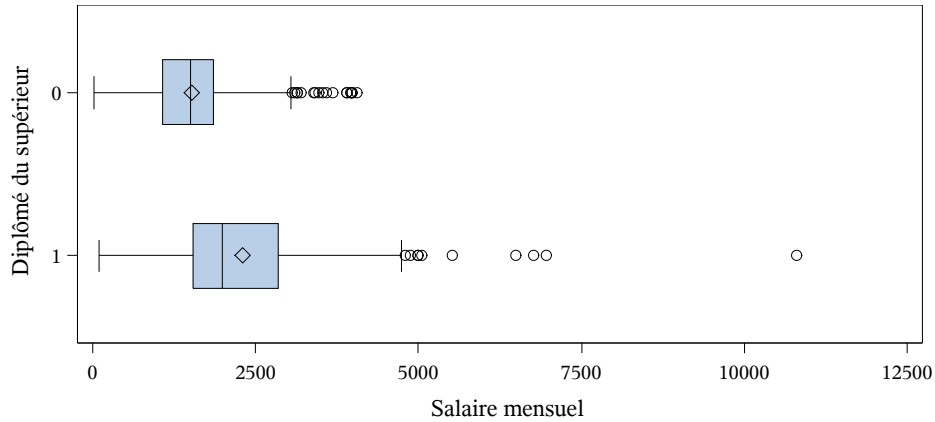


Aucune liaison



Remarque Graphiques réalisés à partir de données simulées.

Boîtes de Tukey : diplôme et salaire




```
PROC SORT DATA = d.ee12t4;
    BY sup;
RUN;
PROC BOXPLOT DATA = d.ee12t4;
    PLOT salred*sup / BOXSTYLE = SCHEMATIC;
RUN; QUIT;
```


Décomposition de la variance

Intuition : **Plus la moyenne de Y varie selon les modalités de X, plus X et Y sont liées**.

La **formule de décomposition de la variance** de Y selon les K modalités de X permet d'approfondir cette intuition :

$$V(Y) = \underbrace{\sum_{k=1}^{K} \frac{n_k}{n-1}(\bar{y_k} - \bar{y})^2}_{\text{Variance inter}} + \underbrace{\sum_{k=1}^{K} \frac{n_k - 1}{n-1} V_k(Y)}_{\text{Variance intra}}$$

où $\bar{y_k}$ et $V_k(Y)$ sont **respectivement la moyenne et la variance de Y quand $X = k$**.

Rapport de corrélation

On définit ainsi le **rapport de corrélation** comme la **part que représente la variance inter dans la variance totale** :

$$\eta^2_{Y|X} = \frac{\text{Variance inter}}{\text{Variance totale}}$$

Propriétés

- $0 \leq \eta^2_{Y|X} \leq 1$ (car $0 \leq$ Variance inter $\leq$ Variance totale)
- plus $\eta^2_{Y|X}$ est proche de 1, plus la liaison entre X et Y est forte.

Remarque Le rapport de corrélation est aussi le R^2 de la régression linéaire de Y sur les modalités de X (*cf.* module 3, 30 janvier-1er février 2017).

Rapport de corrélation : diplôme et salaire

R-Square	Coeff Var	Root MSE	SALRED Mean
0.131892	53.47227	953.9048	1783.924

```
PROC ANOVA DATA = d.ee12t4;
    CLASS sup;
    MODEL salred = sup;
RUN; QUIT;
```

Analyse bivariée : inférence

De l'inférence en statistique bivariée

Dans un cadre bivarié, l'objectif de la statistique inférentielle est de **tester si les relations mises en évidence par les indicateurs de liaison** (*cf.* partie précédente) **sont statistiquement significatives**.

Cela revient à poser un **test statistique**, c'est-à-dire à déterminer s'il est possible de **rejeter une hypothèse nulle** H_0 **au profit d'une hypothèse alternative** H_1.

Si β est un paramètre aléatoire et c une constante, on peut par exemple poser le test :

$$H_0 : \beta = c \quad \text{contre} \quad H_1 : \beta \neq c$$

La **théorie des tests** permet de formaliser d'un point de vue statistique cette prise de décision.

Démarche du test statistique

1. Définition de l'hypothèse nulle H_0.

2. Choix d'une statistique de test de loi connue si H_0 est vraie.
   $\rightarrow$ Quantité dont on sait quelles valeurs elle « devrait » prendre sous H_0.

3. Détermination de la zone de rejet.
   $\rightarrow$ Valeurs de la statistique peu probables si H_0 est vraie.

4. Décision de rejeter ou non l'hypothèse nulle en fonction des observations.

Exemple : Test de dépistage (1)

Les **tests de dépistages** utilisées en laboratoire d'analyse peuvent être envisagés comme des tests statistiques.

En effet, ils ne sont pas totalement déterministes. Pour de multiples raisons, le procédé utilisé pour le dépistage peut **produire des erreurs** :

- soit il peut ne pas détecter un sujet infecté ;
- soit il peut détecter à tort un sujet sain comme infecté.

Le premier type d'erreur est dans ce cas clairement le plus grave : un sujet considéré à tort comme sain ne va pas pouvoir être soigné.

Types d'erreur et définition de l'hypothèse H_0

Comme le test de dépistage, tout test peut produire **deux types d'erreur**.

Test / **Réalité**	H_0 acceptée	H_0 rejetée
H_0 vraie	*Niveau* $1 - \alpha$	Erreur de 1ère espèce α
H_0 fausse	Erreur de 2nd espèce β	*Puissance* $1 - \beta$

Dans ce contexte, la stratégie dite de Neyman-Pearson conduit à **distinguer les deux erreurs selon leur gravité** :

- d'abord l'erreur de première espèce est contrôlée à un seuil α choisi ;
- puis l'erreur de seconde espèce est minimisée autant que possible.

Exemple : Test de dépistage (2)

L'erreur consistant à ne pas détecter des sujets infectés étant la plus grave, dans le cas des tests de dépistage le test statistique est posé de la façon suivante. On teste :

$$H_0 : \{\text{le sujet est infecté}\} \text{ contre } H_1 : \{\text{le sujet est sain}\}$$

L'erreur de première espèce est alors bien {le sujet est infecté mais pas détecté} et l'erreur de seconde espèce {le sujet est sain mais détecté comme infecté}.

On peut alors fixer un niveau aussi élevé que l'on souhaite (99,99 % par exemple) et chercher ensuite à minimiser autant que possible la part de sujets sains détectés comme infectés.

Les erreurs de première espèce couramment acceptées sont 5 % (0,05) et 1 % (0,01). 10 % (0,10) est parfois également acceptée.

Test bilatéral et test unilatéral

La forme de l'hypothèse alternative H_1 détermine si le test est un test dit « bilatéral » ou un test dit « unilatéral » .

Test bilatéral $H_0 : \beta = c$ contre $H_1 : \beta \neq c$

Exemple Test de non-nullité du coefficient de corrélation linéaire.

Test unilatéral $H_0 : \beta = c$ contre $H_1 : \beta > c$

Exemple Test d'indépendance du χ^2 (*cf. infra*).

Choix de la statistique de test

Les propriétés d'un test (en particulier sa puissance à un niveau donné) résultent du **choix de la statistique de test** t.

Remarque On parle également parfois de « statistique pivotale ».

Une statistique de test présente **deux propriétés** :

1. « statistique » : elle est **calculable** à partir des données de l'échantillon ;
2. « de test » : **son comportement sous H_0 est connu** (au moins asymptotiquement)

$$t \hookrightarrow \mathcal{L}(\theta) \quad \text{ou} \quad t \xrightarrow[n \to +\infty]{\mathcal{L}} \mathcal{L}(\theta)$$

avec $\mathcal{L}(\theta)$ une loi tabulée de paramètre θ.

Détermination de la zone de rejet

Intuition : En comparant la valeur prise par la statistique de test aux valeurs qu'elle prend en général quand H_0 est vérifiée, il est possible de juger si H_0 semble une hypothèse raisonnable.

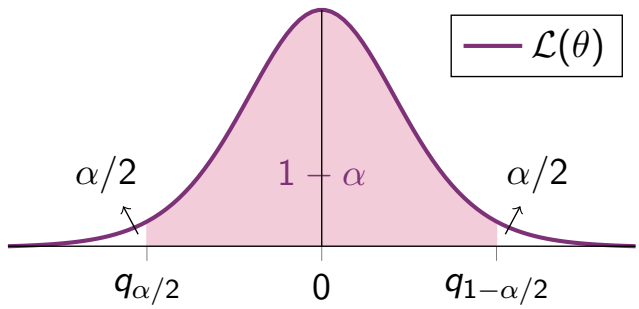
On utilise les quantiles de $\mathcal{L}(\theta)$ pour déterminer une **zone de rejet** W telle que

$$\mathbb{P}(t \in W) = \alpha$$

où α est le risque de première espèce souhaité.

Remarque On parle également de « région critique » pour qualifier la zone de rejet.

Zone de rejet : cas d'un test bilatéral



On sait que, $\mathbb{P}\left(q_{\alpha/2} \leq t \leq q_{1-\alpha/2}\right) = 1-\alpha$ aussi

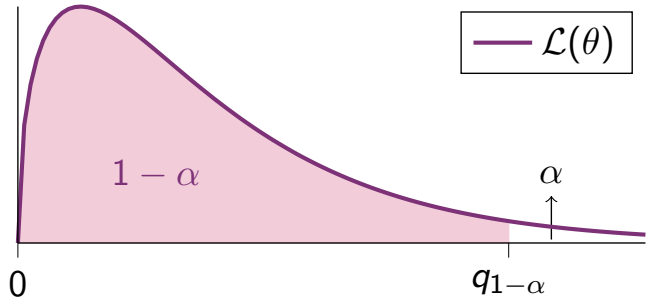
$$\mathbb{P}\left(t < q_{\alpha/2} \text{ ou } t > q_{1-\alpha/2}\right) = \alpha$$

Pour obtenir un test dont l'erreur de première espèce est α, il suffit donc de poser comme région critique :

$$W =]-\infty; q_{\alpha/2}[\cup]q_{1-\alpha/2}; +\infty[$$

On rejette donc l'hypothèse nulle au seuil α quand la statistique de test est inférieure à $q_{\alpha/2}$ ou supérieure à $q_{1-\alpha/2}$.

Zone de rejet : cas d'un test unilatéral



On sait que, $\mathbb{P}(t \leq q_{1-\alpha}) = 1-\alpha$ aussi

$$\mathbb{P}(t > q_{1-\alpha}) = \alpha$$

Pour obtenir un test dont l'erreur de première espèce est α, il suffit donc de poser comme région critique :

$$W =]q_{1-\alpha}; +\infty[$$

On rejette donc l'hypothèse nulle au seuil α quand la statistique de test est supérieure à $q_{1-\alpha}$.

Comparaison à une moyenne théorique (1)

Pour une variable X quelconque de moyenne théorique μ, on souhaite mener le test :

$$H_0 : \mu = m_0 \quad \text{contre} \quad H_1 : \mu \neq m_0$$

avec m_0 une valeur fixée.

En utilisant le théorème central limite, on a montré que :

$$\frac{\bar{X} - \mu}{\hat{\sigma}_X / \sqrt{n}} \xrightarrow[n \to +\infty]{\mathcal{L}} \mathcal{T}_{n-1}$$

Sous l'hypothèse H_0, $\mu = \mathbf{m_0}$ donc la quantité $t = \dfrac{\bar{X} - m_0}{\hat{\sigma}_X / \sqrt{n}}$ (1) est calculable à partir des données (2) suit une loi connue sous H_0 : c'est une **statistique de test**.

Comparaison à une moyenne théorique (2)

Le test est bilatéral donc la zone de rejet est :

$$W =]-\infty; q_{\alpha/2}^{\mathcal{T}_{n-1}}[\cup] q_{1-\alpha/2}^{\mathcal{T}_{n-1}}; +\infty[$$

On rejettera donc l'hypothèse nulle au seuil α quand la statistique de test est inférieure à $q_{\alpha/2}^{\mathcal{T}_{n-1}}$ ou supérieure à $q_{1-\alpha/2}^{\mathcal{T}_{n-1}}$.

Remarque Si on avait posé le test unilatéral

$$H_0 : \mu = m_0 \quad \text{contre} \quad H_1 : \quad \mu > m_0$$

alors la zone de rejet aurait été

$$W =] q_{1-\alpha}^{\mathcal{T}_{n-1}}; +\infty[$$

et on aurait rejeté l'hypothèse nulle au seuil α si la statistique de test avait été supérieure à $q_{1-\alpha}^{\mathcal{T}_{n-1}}$.

Comparaison du salaire moyen à 2 000 € (1)

On cherche à tester si le salaire moyen est égal à 2 000 € au seuil de 1 %. On pose donc le test :

$$H_0 : \mu = 2000 \text{ €} \quad \text{contre} \quad H_1 : \mu \neq 2\,000 \text{ €}$$

La statistique de test associée est

$$t = \frac{\bar{X} - 2\,000}{\hat{\sigma}_X / \sqrt{n}} = \frac{1\,784 - 2\,000}{1\,023/\sqrt{647}} = -5,37$$

Sous l'hypothèse nulle, cette statistique de test suit une loi de *Student* à $n - 1 = 646$ degrés de liberté.

On peut lire dans une table statistique que le quantile à $0,005$ % d'une loi de Student à 646 degrés de liberté est $q_{0,005}^{\mathcal{T}_{646}} = -2,576$.

$t < q_{0,005}^{\mathcal{T}_{646}}$: on peut donc rejeter au seuil de 1 % l'hypothèse que le salaire moyen est égal à 2 000 €.

Comparaison du salaire moyen à 2 000 € (2)

N	Mean	Std Dev	Std Err	Minimum	Maximum
647	1783.9	1023.0	40.2188	24.0000	10798.0

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
1783.9	1704.9	1862.9	1023.0	970.1	1082.0

DF	t Value	Pr > |t|
646	-5.37	<.0001

```
PROC TTEST DATA = d.ee12t4 H0 = 2000;
    VAR salred;
RUN; QUIT;
```


Lecture du test et p-valeur

Pour interpréter un test à partir de la valeur de la statistique de test, il est donc nécessaire de **consulter la table des quantiles** de la loi que suit la statistique de test sous H_0.

Également calculable à partir du test, la **p-valeur** permet de s'affranchir de cette contrainte.

Elle peut être vue comme le **seuil limite au-delà duquel il n'est plus possible de rejeter l'hypothèse nulle**.

Si ce seuil limite est inférieur à un des seuils couramment utilisés (0,10, 0,05, 0,01) alors on peut rejeter l'hypothèse nulle à ce seuil.

Exemple Si la p-valeur d'un test vaut 0,022, alors on peut rejeter l'hypothèse nulle au seuil de 5 % mais pas au seuil de 1 %.

Test d'égalité de deux moyennes indépendantes

On souhaite savoir si la variable X a la même moyenne dans deux groupes 1 et 2 indépendants :

$$H_0 : \mu_1 = \mu_2 \quad \text{contre} \quad H_1 : \mu_1 \neq \mu_2$$

On peut montrer que sous H_0 et sous l'hypothèse d'égalité des variances entre les deux groupes,

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\hat{\sigma}(\frac{1}{n1} + \frac{1}{n_2})} \xrightarrow[n \to +\infty]{\mathcal{L}} \mathcal{T}_{n_1+n_2-2}$$

avec $\hat{\sigma} = \dfrac{(n_1 - 1)\hat{\sigma}_1^2 + (n_2 - 1)\hat{\sigma}_2^2}{n_1 + n_2 - 2}$

Il est donc possible de mener ce test en **comparant la valeur de t aux quantiles d'une loi de Student** à $n_1 + n_2 - 2$ degrés de liberté.

Égalité des salaires selon le diplôme

On cherche à tester l'hypothèse que le salaire moyen est égal selon le niveau de diplôme :

$$H_0 : \mu_{sup=1} = \mu_{sup=0} \quad \text{contre} \quad H_1 : \mu_{sup=1} \neq \mu_{sup=0}$$

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	645	-9.90	<.0001
Satterthwaite	Unequal	287.7	-8.35	<.0001

```
PROC TTEST DATA = d.ee12t4;
    CLASS sup;
    VAR salred;
RUN; QUIT;
```

La statistique de test est inférieure à -1,96 : on peut rejeter l'hypothèse d'égalité au seuil de 5 %. La p-valeur est bien inférieure à 0,05.

Test d'indépendance du χ_2 (1)

Le test d'indépendance entre deux variables qualitatives est **une des applications les plus courantes de la théorie des tests en statistique descriptive**.

Ce test cherche à déterminer si deux variables qualitatives X et Y sont statistiquement liées.

Dans une approche prudente, on pose le test de la façon suivante :

$$H_0 : \{X \text{ et } Y \text{ sont indépendantes}\}$$
$$\text{contre}$$
$$H_1 : \{X \text{ et } Y \text{ ne sont pas indépendantes}\}$$

On contrôle ainsi le risque d'affirmer à tort que X et Y sont liées ($1^{\text{ère}}$ espèce) en minimisant autant que possible le risque de ne pas détecter des associations significatives (2^{nd} espèce).

Test d'indépendance du χ_2 (2)

L'analyse du tableau de contingence a conduit à construire la statistique D^2 (*cf.* partie précédente) :

$$D^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - n_{ij}^*)^2}{n_{ij}^*}$$

Cette quantité est calculable sur les données, et on peut montrer que **sous l'hypothèse H_0 d'indépendance elle suit une loi du χ^2 à $(I-1)(J-1)$ degrés de liberté**.

D^2 est donc une **statistique de test** permettant de mener à bien le test d'indépendance entre X et Y.

Test d'indépendance du χ_2 : Position sur le marché du travail et diplôme (1)

Statistic	DF	Value	Prob
Chi-Square	2	108.0529	<.0001
Likelihood Ratio Chi-Square	2	111.2017	<.0001
Mantel-Haenszel Chi-Square	1	107.6781	<.0001
Phi Coefficient		0.2483	
Contingency Coefficient		0.2410	
Cramer's V		0.2483	

```
PROC FREQ DATA = d.ee12t4;
     TABLES acteu*sup / CHISQ;
     FORMAT acteu $format_acteu.;
RUN;
```


Test d'indépendance du χ_2 : Position sur le marché du travail et diplôme (2)

La statistique de test vaut 108,05 et la valeur critique pour ce test unilatéral au seuil de 5 % est $q^{\chi^2_2}_{0,95} = 5,99$.

$D^2 = 108,05 > 5,99$ donc on peut rejeter l'hypothèse d'indépendance au seuil de 5 %.

Plus encore, la valeur critique à 1 % est $q^{\chi^2_2}_{0,99} = 9,21 < 108,05$: on peut en fait rejeter l'hypothèse d'indépendance au seuil de 1 %.

On aboutit directement à la même conclusion en interprétant la p-valeur, qui est inférieure à 0,01 donc *a fortiori* à 0,05.

Test de significativité de $r_{X,Y}$ (1)

Le test de significativité de $r_{X,Y}$ vise à déterminer si deux variables quantitatives sont statistiquement liées.

En pratique, cela revient à tester si le coefficient de corrélation de Pearson $r_{X,Y}$ est significativement différent de 0.

Dans une approche prudente, on pose le test de la façon suivante :

$$H_0 : r_{X,Y} = 0 \quad \text{contre} \quad H_1 : r_{X,Y} \neq 0$$

On contrôle ainsi le risque d'affirmer à tort que X et Y sont liées ($1^{\text{ère}}$ espèce) en minimisant autant que possible le risque de ne pas détecter des associations significatives (2^{nd} espèce).

Test de significativité de $r_{X,Y}$ (2)

On peut montrer que sous l'hypothèse H_0 de nullité du coefficient de corrélation de Pearson $r_{X,Y}$:

$$t = r_{X,Y} \times \sqrt{\frac{n-2}{1-r_{X,Y}^2}} \hookrightarrow \mathcal{T}_{n-2}$$

En pratique dans SAS, on utilise la p-valeur du test pour accepter ou rejeter l'hypothèse pour un niveau de confiance $1-\alpha$ donné.

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations		
	SALRED	**age**
SALRED Salaire mensuel (EEC 2012T4)	1.00000 647	0.21024 <.0001 647
age Age	0.21024 <.0001 647	1.00000 1752