Statistiques TP 2
Probabilités et statistiques avec Maple

En fait l'agrégation accepte maintenant maple11... donc on on oublie les packages “stats” et “statplots” et on va charger les packages ”Statistics”, (en tapant *with(Statistics):*). ainsi que “plots”.

On va commencer par refaire une partie des mêmes opérations que la semaine dernière avec le nouveau package : ainsi on travaille au début sur une variable aléatoire supposée modéliser la taille d'une population, qui suit une loi normale $N(\mu, \sigma^2)$ avec $\mu = 175$ et $\sigma = 10$.

1 Manipulation de lois, utilisation de commandes de base

On va commencer par définir X ainsi :
X:=RandomVariable(Normal(175,10));
(le deuxième paramètre est là encore l'écart-type, pas la variance...) La commande RandomVariable permet de définir toutes sortes de loi, discrètes ou à densité. Faire **?Distributions** pour en avoir une liste.

1. La densité de X s'obtient à l'aide de la commande $\mathbf{PDF}(\mathbf{X}, \mathbf{x})$. Obtenir la valeur de la densité en 180. Pour obtenir l'espérance de X, on peut utiliser $\mathbf{Mean}(\mathbf{X})$, la variance à l'aide de $\mathbf{Variance}(\mathbf{X})$, et l'espérance de X^3 à l'aide de la commande $\mathbf{ExpectedValue}(\mathbf{X^3})$.
2. Tracer la densité avec la commande $\mathbf{DensityPlot}(\mathbf{X})$.
3. Déterminer le quantile à $97,5\%$ de X, en utilisant la commande $\mathbf{Quantile}(\mathbf{X}, \mathbf{0.975})$.
4. Quelle est la loi de $\frac{X-175}{10}$? Vérifier en obtenant sa densité l'aide de $\mathbf{PDF}((\mathbf{X} - \mathbf{175})/\mathbf{10})$.
5. Si on définit $\mathbf{Y} := \mathbf{RandomVariable}(\mathbf{175}, \mathbf{10})$, maple considère que c'est une variable indépendante de X. C'est donc différent que de poser $U := X$. Obtenir par exemple la densité de $X + Y$, et comparer avec la densité de $X + U$.

2 Statistiques sur un échantillon de taille n

En statistique, on ne connaît en général pas les détails de la loi qu'on observe, mais on a accès à des observations d'occurrence de cette loi, qu'on suppose indépendantes. On appelle cela un n-échantillon de loi X.

1. Générer un 100-échantillon *Ech* de loi normale $N(175, 100)$ à l'aide de la commande $\mathbf{Ech} := \mathbf{Sample}(\mathbf{X}, \mathbf{100})$. Un tel échantillon est un vecteur (un type un peu particulier, mais qui se comporte essentiellement comme un tableau, et pas comme une liste).
2. Calculer la moyenne empirique $\bar{x} = \frac{\sum_{i=1}^{n} X_i}{n}$ de l'échantillon à l'aide de la commande **Mean**.
3. Calculer l'écart-type empirique $\bar{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{x})^2}{n-1}}$ (attention au $n-1$, on verra sa justification plus tard en TD...) à l'aide de **StandardDeviation**.
4. Calculer le quantile correspondant à 0.975 à l'aide de **Quantile**. Ce sont donc les mêmes commandes pour les valeurs théoriques pour X et pour les valeurs empiriques mesurées sur *Ech*. Elles acceptent comme paramètre un échantillon qui peut être un vecteur ou une liste.
5. Tracer un histogramme correspondant à l'échantillon *Ech* à l'aide de la commande $Histogram$. L'option *bincount* permet de contrôler le nombre de barres verticales. Le représenter sur le même graphique que la vraie densité de X et que la densité estimée $N(\bar{x}, \bar{\sigma})$.

6. Tracer la fonction de répartition empirique en traçant la suite de points de coordonnées $(Quantile(i/100), i/100)$.

   Rappel pour tracer une suite de points de coordonnées A_i, B_i on peut utiliser
   plot([**seq**([$\mathbf{A_i}, \mathbf{B_i}$], $\mathbf{i = 1..100}$)]);

3 Illustration de la loi des grands nombres

Rappel : La loi des grands nombres est le théorème qui dit que si $(X_n)_{n\in N}$ est une suite de variables aléatoires indépendantes intégrables, de même loi, d'espérance m, alors pour presque tout ω, $\frac{\sum_{i=1}^{n} X_i(\omega)}{n}$ converge vers m : la moyenne arithmétique converge vers l'espérance. Cela entraine également ici presque sûrement la convergence des fonctions de répartition empiriques vers la fonction de répartition théorique.

On va par exemple travailler avec une variable aléatoire X de loi uniforme sur $\{1, 2, 3, 4, 5, 6\}$ (on va simuler un lancer de dés). On peut donc commencer par poser $X := DiscreteUniform(1, 6)$.

1. Quelle est l'espérance m de X ? Quel est son écart-type σ?

2. Générer un échantillon de X de taille 5000; représenter pour n allant de 30 à 5000 la moyenne empirique $\bar{x}$ des n premiers termes en fonction de n (le plus simple est d'utiliser la commande **CumulativeSum**(**Ech**)) ainsi qu'une droite horizontale correspondant à l'espérance.

3. Représenter les fonctions de répartition empiriques pour les 100, 500,1000 et 5000 premiers termes ainsi que la fonction de répartition théorique. Pour cela, vous pouvez définir 3 listes comportant respectivement les 100, 500 et 1000 premiers termes de votre échantillon. Vous pouvez pour cela soit effectuer une boucle pour chaque liste, soit convertir votre échantillon en liste (à l'aide de la commande **convert**(**Ech**, **list**) et utliser le fait que si l est une liste de N termes, **l2** := **l**[**1..p**] fournit une liste avec les p premiers termes. Pour la fonction de répartition théorique, on peut utiliser $CDF(X, x)$ qui fournit la valeur de la fonction de répartition en x.

4. On considère maintenant une variable aléatoire suivant une loi de Cauchy de paramètres $(0, 1)$ $Y := RandomVariable(Cauchy(0, 1))$;. Quelle est la densité de Y ? Y est-elle d'espérance finie ? Effectuer les mêmes opérations que sur X.

4 Théorème central limite

1. On l'a vu, la moyenne empirique converge vers l'espérance. Représenter pour des tailles n allant de 30 à 5000 $\sqrt{n}\sigma(\bar{x} - m)$.

   Cette quantité ne semble pas converger, mais elle reste d'ordre 1. Néanmoins, le théorème central limite affirme que si (X_n) est une suite de variables aléatoires indépendantes de même loi et d'espérance μ et de variance σ^2 finies, $\sqrt{n}\sigma(\bar{x} - \mu)$ converge en loi vers une loi normale centrée réduite. Cette quantité reste donc aléatoire, mais on sait comment elle est distribuée asymptotiquement.

2. On va commencer par observer ce phénomène sur une loi binomiale. On peut en fait l'observer sur les graphes des lois : soit Z une variable aléatoire qui suit une loi binomiale $B(n, p)$ ($Z := Binomial(n, p)$).

   (a) Rappeler l'espérance de Z et sa variance. Pour $n = 500$ et $p = 0.3$, obtenir $P(Z = 40)$ à l'aide de la commande $ProbabilityFunction$. Obtenir le graphe des $(k, P(Z = k))$ à l'aide de $DensityPlot$. Comparer avec la densité d'une variable aléatoire normale de même espérance et variance.

(b) Vers quoi converge en loi $\frac{Z-np}{\sqrt{np(1-p)}}$? Faire une procédure qui prend les paramètres de la loi binomiale n et p en entrée, et trace le graphe formé des points $(M_k)_{0\leq k\leq n}$ de coordonnées $(\frac{k-np}{\sqrt{np(1-p)}}, \sqrt{np(1-p)}P(Z=k))$, ainsi que la densité de la loi normale centrée réduite.
Rappel : une procédure se présente sous la forme :
$Bidule := proc(parametres)$
$instructions$;
$end\,proc$;
Pour aller à la ligne sans effectuer tout de suite l'instruction avec maple, faire "shift+ entrée". Tester la procédure avec diverses valeurs de n et p. (prendre $n \leq 1000$).

(c) On peut aussi voir le théorème central limite sur des échantillons. Faire une procédure qui prend n, p et N_{obs} en entrée, génère un N_{obs} échantillon $(Z_1, \dots Z_{Nobs})$ de loi binomiale $B(n,p)$, obtient un vecteur comportant les $\frac{Z_i-np}{\sqrt{np(1-p)}}$ (en utilisant $\mathbf{map(x->f(x),v)}$ qui applique la fonction f à tous les éléments d'une liste ou d'un vecteur), trace un histogramme des $\frac{Z_i-np}{\sqrt{np(1-p)}}$ ainsi que la densité de la loi normale centrée réduite. Lancer la procédure avec $N_{obs} = 1000$, $n = 100$ et $p = 0.1$ puis 0.3, 0.5, 0.99. Recommencer avec $n = 300$.

3. On reprend le lancer de dé de tout à l'heure, et on va utiliser la fonction de répartition, plus fiable que l'histogramme. Construire une procédure qui prend en entrée deux entiers n et N_{obs}, génère N_{obs} n-échantillons de X, calcule pour chacun de ces échantillons $\frac{\sqrt{n}}{\sigma}(\bar{x}-m)$, stocke ces valeurs dans une liste, et enfin trace sur un même graphique la fonction de répartition empirique de ces grandeurs ainsi que le graphe de la fonction de répartition empirique de la loi normale centrée réduite.

4. Tester la procédure avec $Nobs = 1000$ et $n = 30, 50, 100, 300, 1000$.