14.1 Soit X une VAR d'espérance m et une variance v.
On dispose d'un n-échantillon $(X_1, \dots, X_n)$ de X.
On appelle variance empirique de X la variable $W_n = \dfrac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X_n}^2$, où $\overline{X_n}$ est la moyenne empirique de X.

1. Calculer $\mathbb{E}\big[\overline{X_n}\big]$ et $\mathbb{V}\big[\overline{X_n}\big]$, et en déduire $\mathbb{E}\big[\overline{X_n}^2\big]$.
2. Calculer $\mathbb{E}\big[W_n\big]$ et en déduire un estimateur sans biais de v.

1. La moyenne empirique $\overline{X_n}$ est définie par
$$\overline{X_n} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$
On a donc par linéarité de l'espérance :
$$\mathbb{E}[\overline{X_n}] = \mathbb{E}\left[\frac{1}{n}(X_1 + X_2 + \cdots + X_n)\right] = \frac{1}{n}(\mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n]) = \frac{1}{n}nm = \boxed{m}$$
et par indépendance des variables $X_1, \dots, X_n$ (puisque c'est un échantillon de X) :
$$\mathbb{V}[\overline{X_n}] = \mathbb{V}\left[\frac{1}{n}(X_1 + X_2 + \cdots + X_n)\right] = \frac{1}{n^2}(\mathbb{V}[X_1] + \cdots + \mathbb{V}[X_n]) = \frac{1}{n^2}nv = \boxed{\frac{v}{n}}$$
Or, on sait que $\mathbb{V}[\overline{X_n}] = \mathbb{E}[\overline{X_n}^2] - (\mathbb{E}[X_n])^2$. On en déduit que
$$\mathbb{E}[\overline{X_n}^2] = \mathbb{V}[\overline{X_n}] + (\mathbb{E}[X_n])^2 = \frac{v}{n} + m^2$$

2. On a donc :
$$\begin{aligned}
\mathbb{E}[X_n] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X_n}^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i^2] - \mathbb{E}[\overline{X_n}^2] \\
&= \mathbb{E}[X_1^2] - \left(\frac{v}{n} + m^2\right) \\
&= \left(\mathbb{V}[X_1] + (\mathbb{E}[X_1])^2\right) - \frac{v}{n} - m^2 \\
&= v - \frac{v}{n} \\
&= \boxed{\frac{n-1}{n}v}
\end{aligned}$$
On remarque que la variance empirique W_n n'est pas un estimateur sans biais de v (puisqu'on a $\mathbb{E}[W_n] \neq v$), mais, puisque $\lim\limits_{n\to+\infty} \dfrac{n-1}{n} = 1$, W_n est un estimateur asymptotiquement sans biais de v.
Pour obtenir un estimateur sans biais de v, on pose $W'_n = \dfrac{n}{n-1}W_n$:
$$\mathbb{E}[W'_n] = \mathbb{E}\left[\frac{n}{n-1}W_n\right] = \frac{n}{n-1}\mathbb{E}[W_n] = \frac{n}{n-1}\frac{n-1}{n}v = v$$
donc W'_n est, lui, un estimateur sans biais de v.

14.2 Soit X une VAR de loi uniforme sur un intervalle $[0,a]$ où a est un paramètre inconnu, et on dispose de $(X_1,\dots,X_n)$ un n-échantillon de X. On note $\overline{X_n}$ la moyenne empirique de X.

1. Soit $T_n = 2\overline{X_n}$. Montrer que T_n est un estimateur sans biais de a et calculer son risque quadratique.
2. Soit $T'_n = \max(X_1,\dots,X_n)$. Donner la fonction de répartition de T'_n. En déduire une densité de T'_n, puis son biais et son risque quadratique.
3. Soit $T''_n = \dfrac{n+1}{n}T'_n$. Déterminer son biais et son risque quadratique.
4. Pour de grandes valeurs de n, quelle est le meilleur estimateur de a ?

X est une VAR de loi uniforme sur $[0,a]$. Rappelons donc que sa fonction de répartition est :

$$\forall x\in\mathbb{R},\ F(x) = \begin{cases} 0 & \text{si } x\leqslant 0 \\ \dfrac{x}{a} & \text{si } x\in[0,a] \\ 1 & \text{si } x\geqslant a \end{cases}$$

et qu'on a $\mathbb{E}[X] = \frac{a}{2}$ et $\mathbb{V}[X] = \frac{a^2}{12}$

1.

$$\begin{aligned}\mathbb{E}[T_n] = \mathbb{E}[2\overline{X_n}] &= 2\mathbb{E}[\overline{X_n}] \\ &= 2\mathbb{E}\left[\frac{1}{n}(X_1+\cdots+X_n)\right] \\ &= \frac{2}{n}\left(\mathbb{E}[X_1]+\cdots+\mathbb{E}[X_n]\right) \\ &= \frac{2}{n}\left(n\frac{a}{2}\right) \\ &= a\end{aligned}$$

donc T_n est bien un estimateur sans biais de a.
Puisque T_n est un estimateur sans biais de a, son risque quadratique est égal à sa variance.

$$\begin{aligned}r(T_n) &= \mathbb{V}[T_n] + (b(T_n))^2 \\ &= \mathbb{V}[T_n] \\ &= \mathbb{V}[2\overline{X_n}] \\ &= 4\mathbb{V}\left[\frac{1}{n}(X_1+\cdots+X_n)\right] \\ &= \frac{4}{n^2}\left(\mathbb{V}[X_1]+\cdots\mathbb{V}[X_n]\right) \\ &= \frac{4}{n^2}\left(n\frac{a^2}{12}\right) \\ &= \frac{a^2}{3n}\end{aligned}$$

2. On a $X_1(\Omega) = \cdots = X_n(\Omega) = [0,a]$, donc $T'_n(\Omega) = [0,a]$ également.

Notons G la fonction de répartition de T'_n.
- $\forall x<0$, $G(x) = \mathbb{P}(T'_n\leqslant x) = 0$
- $\forall x\geqslant a$, $G(x) = \mathbb{P}(T'_n\leqslant x) = 1$

- Soit $x \in [0, a[$, alors

$$\begin{aligned} G(x) = \mathbb{P}(T'_n \leqslant x) &= \mathbb{P}\left(\max(X_1, \dots, X_n) \leqslant x\right) \\ &= \mathbb{P}\left([X_1 \leqslant x] \cap [X_2 \leqslant x] \cap \dots \cap [X_n \leqslant x]\right) \\ &= \mathbb{P}(X_1 \leqslant x)\mathbb{P}(X_2 \leqslant x) \cdots \mathbb{P}(X_n \leqslant x) \quad \text{(indépendance des } X_i\text{)} \\ &= (F(x))^n \\ &= \frac{x^n}{a^n} \end{aligned}$$

On a donc

$$\forall x \in \mathbb{R},\ G(x) = \begin{cases} 0 & \text{si } x \leqslant 0 \\ \dfrac{x^n}{a^n} & \text{si } x \in [0, a] \\ 1 & \text{si } x \geqslant a \end{cases}$$

On remarque que la fonction G est clairement continue sur $\mathbb{R}$ et de classe $\mathcal{C}^1$ sur au moins $\mathbb{R} \setminus \{0, a\}$, ce qui prouve que la variable T'_n est bien une variable aléatoire à densité. Pour obtenir une densité de T'_n, il nous suffit de dériver G là où cela est possible et de compléter par des valeurs positives ailleurs. Une densité de T'_n est donc par exemple :

$$\forall x \in \mathbb{R},\ g(x) = \begin{cases} 0 & \text{si } x < 0 \\ \dfrac{nx^{n-1}}{a^n} & \text{si } x \in [0, a] \\ 0 & \text{si } x > a \end{cases}$$

Calculons le biais de T'_n :

$$\begin{aligned} \mathbb{E}[T'_n] &= \int_{-\infty}^{+\infty} t g(t) dt \\ &= \int_0^a t g(t) dt \\ &= \frac{n}{a^n} \int_0^a t^n dt \\ &= \frac{n}{a^n} \left[\frac{t^{n+1}}{n+1}\right]_0^a \\ &= \frac{n}{n+1} a \end{aligned}$$

Puisque $\mathbb{E}[T'_n] \neq a$, T'_n est un estimateur biaisé de a et on a :

$$b(T'_n) = \mathbb{E}[T'_n] - a = \frac{n}{n+1}a - a = \boxed{a\left(\frac{-1}{n+1}\right)}$$

Calculons la variance de T'_n.

$$\mathbb{E}[(T'_n)^2] = \int_{-\infty}^{+\infty} t^2 g(t) dt = \frac{n}{a^n} \int_0^a t^{n+1} dt = \frac{n}{a^n} \left[\frac{t^{n+2}}{n+2}\right]_0^a = \frac{n}{n+2} a^2$$

On en déduit que

$$\mathbb{V}[T'_n] = \mathbb{E}[(T'_n)^2] - \mathbb{E}[T'_n]^2 = \frac{n}{n+2}a^2 - \frac{n^2}{(n+1)^2}a^2 = a^2 \frac{n}{(n+2)(n+1)^2}$$

Ainsi :

$$\begin{aligned} r(T'_n) &= \mathbb{V}[T'_n] + b(T'_n)^2 \\ &= a^2\frac{n}{(n+2)(n+1)^2} + \frac{a^2}{(n+1)^2} \\ &= \boxed{a^2\frac{2}{(n+1)(n+2)}} \end{aligned}$$

3. $$\mathbb{E}[T''_n] = \mathbb{E}\left[\frac{n+1}{n}T'_n\right] = \frac{n+1}{n}\mathbb{E}[T'_n] = \frac{n+1}{n}\frac{n}{n+1}a = a$$

   donc T''_n est un estimateur sans biais de a. Son risque quadratique est donc égal à sa variance

$$r(T''_n) = \mathbb{V}[T''_n] = \mathbb{V}\left[\frac{n+1}{n}T'_n\right] = \frac{(n+1)^2}{n^2}\mathbb{V}[T'_n] = a^2\frac{1}{n(n+2)}$$

4. Récapitulons les informations pour nos estimateurs pour n assez grand :
   - T_n : $b(T_n) = 0$ et $r(T_n) \underset{n\to+\infty}{\sim} \dfrac{a^2}{3n}$
   - T'_n : $b(T'_n) \underset{n\to+\infty}{\sim} -\dfrac{a}{n}$ et $r(T'_n) \underset{n\to+\infty}{\sim} \dfrac{2a^2}{n^2}$
   - T''_n : $b(T''_n) = 0$ et $r(T''_n) \underset{n\to+\infty}{\sim} \dfrac{a^2}{n^2}$

   T_n et T''_n sont des estimateurs sans biais, mais T'_n est asymptotiquement sans biais, donc lorsque n devient grand, les trois sont "à peu près" sans biais. Il s'agit de comparer les risques quadratiques.
   T''_n est l'estimateur le plus performant : sans biais et risque quadratique très petit
   Entre T'_n et T_n, T_n est préférable si n est petit (car sans biais), mais T'_n devient meilleur si n devient grand car son risque quadratique est plus petit.
   On a donc en terme de performance :

$$T''_n \gg T'_n \gg T_n$$

14.3 Lors d'un sondage sur 100 personnes interrogées, 60 pensent voter pour A.

On modélise ce résultat par un échantillon $(X_1, X_2, \ldots, X_{100})$ de variables indépendantes de même loi de Bernoulli de paramètre p.
On cherche à déterminer un intervalle de confiance pour p au niveau de confiance 99%.

1. Déterminer l'espérance et la variance de la moyenne empirique $F = \dfrac{1}{100}\displaystyle\sum_{i=1}^{100} X_i$
2. On note F^* la variable centrée réduite associée à F. Par quelle loi peut-on approcher celle de F^* ?
   Déterminer t tel que $\mathbb{P}(-t \leqslant F^* \leqslant t) \geqslant 0.99$ et en déduire que

$$\mathbb{P}\left(F - t\frac{\sqrt{p(1-p)}}{10} \leqslant p \leqslant F + t\frac{\sqrt{p(1-p)}}{10}\right) \geqslant 0.99$$

3. Montrer que pour tout $p \in [0,1]$, $p(1-p) \leqslant \dfrac{1}{4}$ et en déduire un intervalle de confiance pour p au niveau de confiance 0.99, puis en donner une estimation.

1. Les X_i suivent toutes des loi de Bernoulli de paramètre p. On a donc pour tout i, $\mathbb{E}[X_i] = p$ et $\mathbb{V}[X_i] = p(1-p)$. On a donc par linéarité de l'espérance,

$$\mathbb{E}[F] = \mathbb{E}\left[\frac{1}{100}\sum_{i=1}^{100} X_i\right] = \frac{1}{100}\sum_{i=1}^{100}\mathbb{E}[X_i] = \frac{1}{100}100p = p$$

De plus, par indépendance des variables $X_1, \ldots, X_{100}$,

$$\mathbb{V}[F] = \mathbb{V}\left[\frac{1}{100}\sum_{i=1}^{100} X_i\right] = \frac{1}{100^2}\sum_{i=1}^{100}\mathbb{V}[X_i] = \frac{1}{100^2}100p(1-p) = \frac{p(1-p)}{100}$$

2. Les X_i sont toutes indépendantes, identiquement distribuées de loi de Bernoulli de paramètre p, donc d'après le Théorème de la Limite Centrée, puisqu'on somme 100 variables indépendantes ($100 > 30$), on sait qu'on peut approcher la loi de $F^* = \dfrac{F - \mathbb{E}[F]}{\sqrt{\mathbb{V}[F]}}$ par une loi normale centrée réduite $\mathcal{N}(0,1)$.
On pourra donc à présent supposer que $F^* \rightsquigarrow \mathcal{N}(0,1)$.

$$\mathbb{P}(-t \leqslant F^* \leqslant t) \geqslant 0.99 \iff 2\Phi(t) - 1 \geqslant 0.99 \iff \Phi(t) \geqslant 1.99 \iff \Phi(t) \geqslant 0.995 \iff t \geqslant 2.58$$

On peut donc choisir $\boxed{t = 2.58}$.
On a pour le t précédent :

$$\begin{aligned}
\mathbb{P}(-t \leqslant F^* \leqslant t) \geqslant 0.99 &\iff \mathbb{P}\left(-t \leqslant \frac{F-p}{\sqrt{\frac{p(1-p)}{100}}} \leqslant t\right) \geqslant 0.99 \\
&\iff \mathbb{P}\left(-t\frac{\sqrt{p(1-p)}}{10} \leqslant F - p \leqslant t\frac{\sqrt{p(1-p)}}{10}\right) \geqslant 0.99 \\
&\iff \mathbb{P}\left(-t\frac{\sqrt{p(1-p)}}{10} \leqslant p - F \leqslant t\frac{\sqrt{p(1-p)}}{10}\right) \geqslant 0.99 \\
&\iff \mathbb{P}\left(F - t\frac{\sqrt{p(1-p)}}{10} \leqslant p \leqslant F + t\frac{\sqrt{p(1-p)}}{10}\right) \geqslant 0.99
\end{aligned}$$

3. Etudions la fonction $g : x \mapsto x(1-x)$ sur $[0,1]$
g est dérivable sur $[0,1]$ et $\forall x \in [0,1]$, $g'(x) = 1 - 2x$. La fonction est croissante sur $[0, 1/2]$ et décroissante sur $[1/2, 1]$. Ainsi, pour tout $x \in [0,1]$, $g(x) \leqslant g(1/2) = \dfrac{1}{4}$
On en déduit que pour tout $p \in [0,1]$, $\sqrt{p(1-p)} \leqslant \dfrac{1}{2}$. Quitte à élargir un peu l'intervalle de confiance obtenu précédemment, on peut donc affirmer que

$$\mathbb{P}\left(F - \frac{t}{20} \leqslant p \leqslant F + \frac{t}{20}\right) \geqslant 0.99$$

(puisque l'événement $[F - \dfrac{t}{20} \leqslant p \leqslant F + \dfrac{t}{20}]$ est inclus dans $[F - t\dfrac{\sqrt{p(1-p)}}{10} \leqslant p \leqslant F + t\dfrac{\sqrt{p(1-p)}}{10}]$).

L'intervalle de confiance à 99% pour p est donc $\left[F - \frac{2.58}{20}, F + \frac{2.58}{20}\right] = [F - 0.129, F + 0.129]$. Puisque l'énoncé nous donne une réalisation de F : 0.6, une bonne estimation de cet intervalle de confiance est :

$$[0.471, 0.729]$$

14.4 Afin d'étudier la proportion p de consommateurs satisfaits par un produit, on a interrogé 100 consommateurs. 56 d'entre eux ont déclaré être satisfaits par le produit. Donner un intervalle de confiance à 95% de p.

On applique la même méthode que précédemment.

On a interrogé les 100 consommateurs. On note leurs résultats $X_1, \dots, X_{100}$ sous la forme $X_i = 1$ si le i-ième consommateur est satisfait, et $X_i = 0$ si le i-ième consommateur est mécontent du produit.

Les X_i suivent tous une loi de Bernoulli de paramètre p, et sont indépendantes.

On note $Y = \dfrac{X_1 + X_2 + \cdots + X_{100}}{100}$. On a $\mathbb{E}[Y] = p$ et $\mathbb{V}[Y] = \dfrac{p(1-p)}{n}$

Le théorème de la Limite Centrée nous dit que $\dfrac{Y - \mathbb{E}[Y]}{\sqrt{\mathbb{V}[Y]}} = 10\dfrac{Y-p}{\sqrt{p(1-p)}}$ suit approximativement une loi normale centrée réduite $\mathcal{N}(0,1)$ puisque on a un nombre assez grand de variables indépendantes.

Soit Z une loi normale centrée réduite. Cherchons un $t > 0$ tel que

$$\mathbb{P}\left(-t \leqslant Z \leqslant t\right) \geqslant 0.95$$

i.e.

$$2\Phi(t) - 1 \geqslant 0.95 \iff 2\Phi(t) \geqslant 1.95 \iff \Phi(t) \geqslant 0.975 \iff t \geqslant 1.96$$

On peut donc prendre $\underline{t = 1.96}$.

Alors :

$$\begin{aligned}\mathbb{P}\left(-t \leqslant Z \leqslant t\right) \geqslant 0.95 &\iff \mathbb{P}\left(-t \leqslant 10\frac{Y-p}{\sqrt{p(1-p)}} \leqslant t\right) \geqslant 0.95 \\ &\iff \mathbb{P}\left(Y - t\frac{\sqrt{p(1-p)}}{10} \leqslant p \leqslant Y + t\frac{\sqrt{p(1-p)}}{10}\right) \geqslant 0.95\end{aligned}$$

Donc quitte à élargir encore l'intervalle de confiance, puisque $\forall p \in [0,1]$, $p(1-p) \leqslant \dfrac{1}{4}$ et $\sqrt{p(1-p)} \leqslant \dfrac{1}{2}$, on en déduit que

$$\mathbb{P}\left(Y - \frac{t}{20} \leqslant p \leqslant Y + \frac{t}{20}\right) \geqslant 0.95$$

L'intervalle de confiance à 95% pour p est donc

$$\left[Y - \frac{1.96}{20}, F + \frac{1.96}{20}\right] = [Y - 0.098, F + 0.098]$$

Puisque l'énoncé nous donne une réalisation de Y : 0.56, une bonne estimation de cet intervalle de confiance est :

$$[0.56 - 0.098, 0.56 + 0.098] = [0.462, 0.658]$$

14.5 Dans un scrutin, le dépouillement des n premiers bulletins donne 60% de votes favorables au candidat A.

1. Déterminer n pour que l'on puisse affirmer avec moins de 5% de risque d'erreurs que A obtiendra entre 58% et 62% de voix.
2. On suppose que A obtient effectivement 60% de voix. Trouver la probabilité pour que les partisans de A soient en minorité dans un échantillon donné de 100 électeurs.

1. Il nous faut trouver $[0.58, 0.62]$ comme intervalle de confiance au niveau de confiance 95% pour la proportion p de bulletins favorables à A.
   Cet intervalle de confiance est à priori de la forme :
   $$\left[x_0 - t\frac{1}{2\sqrt{n}}, x_0 + t\frac{1}{2\sqrt{n}}\right]$$
   avec x_0 l'estimateur ponctuel de p (ici $x_0 = 0.6$) et $t > 0$ tel que $2\Phi(t) - 1 \geqslant 0.95$, donc $t = 1.96$ convient.
   L'intervalle de confiance est donc
   $$\left[0.6 - \frac{1.96}{2\sqrt{n}}, 0.6 + \frac{1.96}{2\sqrt{n}}\right]$$
   Finalement, tout intervalle de ce type avec $\dfrac{1.96}{2\sqrt{n}} \leqslant 0.02$ convient.
   On choisit alors n tel que $n \geqslant \left(\dfrac{1.96}{2 \times 0.02}\right)^2 = 2401$.
   On peut donc choisir $\underline{n = 2401}$.
2. Notons X le nombre de partisans de A dans un échantillon donné de 100 électeurs. La variable X suit une loi binomiale $\mathcal{B}(100, 0.6)$. On a $\mathbb{E}[X] = 100 \times 0.6 = 60$ et $\mathbb{V}[X] = 100 \times 0.6 \times 0.4 = 24$.
   Puisque le premier paramètre de la loi binomiale est assez élevé ($100 > 30$), on sait qu'on peut approcher X par une loi normale, et plus particulièrement $\dfrac{X - 60}{\sqrt{24}}$ suit approximativement une loi normale centrée réduite $\mathcal{N}(0, 1)$.
   $$\mathbb{P}(X < 50) = \mathbb{P}(X - 60 < -10) = \mathbb{P}\left(\frac{X - 60}{\sqrt{24}} < \frac{-10}{\sqrt{24}}\right) \simeq \Phi\left(-\frac{10}{\sqrt{24}}\right) = 1 - \Phi\left(\frac{10}{\sqrt{24}}\right)$$
   Puisque $\dfrac{10}{\sqrt{24}} \simeq 2.04$, on a $1 - \Phi(2.04) \simeq 0.0217$ d'après les tables de la loi normale.
   La probabilité pour que les partisans de A soient en minorité est donc d'environ 0.0217.

14.6 Une usine fabrique des câbles. On suppose que la charge maximale supportée par un câble exprimée en tonnes est une variable aléatoire suivant une loi normale de paramètres m et 0.5.
Une étude portant sur 50 câbles a donné une moyenne des charges maximales supportées égale à 12.2 tonnes.

1. Déterminer l'intervalle de confiance à 99% de la charge maximale moyenne de tous les câbles fabriqués par l'usine.
2. Quelle doit être la taille minimale de l'échantillon étudié pour que la longueur de l'intervalle de confiance à 99% soit inférieure ou égale à 0.2 ?

1. On veut un intervalle de confiance à 99%.
   Soit Z une variable aléatoire centrée réduite. Déterminons un t tel que

$$\mathbb{P}(-t \leqslant Z \leqslant t) \geqslant 0.99 \iff 2\Phi(t) - 1 \geqslant 0.99 \iff \Phi(t) \geqslant 0.995 \iff t \geqslant 2.58$$

   On peut prendre $t = 2.58$.
   L'intervalle de confiance de m à un niveau de confiance 99% est

$$\left[x_0 - t\frac{1}{2\sqrt{50}}, x_0 + t\frac{1}{2\sqrt{50}}\right]$$

   où x_0 est l'estimateur ponctuel de m, ici 12.2
   Ici, l'intervalle de confiance est :

$$\left[12.2 - 2.58\frac{1}{2\sqrt{50}}, 12.2 + 2.58\frac{1}{2\sqrt{50}}\right] = [12.0175, 12.3824]$$

2. Si on a un intervalle de confiance à 99% avec un échantillon de n objets, il est de la forme

$$12.2 - 2.58\frac{1}{2\sqrt{n}}, 12.2 + 2.58\frac{1}{2\sqrt{n}}$$

   La longueur de l'intervalle de confiance est $2 \times 2.58\dfrac{1}{2\sqrt{n}} = \dfrac{2.58}{\sqrt{n}}$
   On veut donc que

$$\frac{2.58}{\sqrt{n}} \leqslant 0.2 \iff \left(\frac{2.58}{0.2}\right) \leqslant \sqrt{n} \iff n \geqslant \left(\frac{2.58}{0.2}\right)^2 \simeq 366.41$$

   On peut prendre par exemple $n = 367$ comme taille minimale de l'échantillon pour que la longueur de l'intervalle de confiance à 99% soit inférieure à 0.2.