

Probabilités, Statistiques, Combinatoire

Philippe Duchon

Université de Bordeaux – Licence Informatique

2017-2018

Statistiques

- ▶ En **probabilités**, on modélise une situation, et on cherche à faire des calculs qui servent à faire des *prédictions* (on calcule la probabilité que tel événement se produise, qu'on observe tel phénomène)

Statistiques

- ▶ En **probabilités**, on modélise une situation, et on cherche à faire des calculs qui servent à faire des *prédictions* (on calcule la probabilité que tel événement se produise, qu'on observe tel phénomène)
- ▶ En **statistiques**, la démarche est inverse : on part d'observations ("échantillon"), et on cherche à extraire un modèle, ou à répondre à une question

Statistiques

- ▶ En **probabilités**, on modélise une situation, et on cherche à faire des calculs qui servent à faire des *prédictions* (on calcule la probabilité que tel événement se produise, qu'on observe tel phénomène)
- ▶ En **statistiques**, la démarche est inverse : on part d'observations ("échantillon"), et on cherche à extraire un modèle, ou à répondre à une question
- ▶ (Évaluation de la qualité d'un lot dans une production ; estimation de l'effet d'un traitement par rapport à l'absence de traitement ; sondages ; ...)

Statistiques

- ▶ En **probabilités**, on modélise une situation, et on cherche à faire des calculs qui servent à faire des *prédictions* (on calcule la probabilité que tel événement se produise, qu'on observe tel phénomène)
- ▶ En **statistiques**, la démarche est inverse : on part d'observations ("échantillon"), et on cherche à extraire un modèle, ou à répondre à une question
- ▶ (Évaluation de la qualité d'un lot dans une production ; estimation de l'effet d'un traitement par rapport à l'absence de traitement ; sondages ; ...)
- ▶ Par nature, la "décision" qu'on est amené à prendre est toujours entachée d'un risque d'erreur ; l'objectif de la théorie statistique est de déterminer les "bonnes" procédures pour contrôler ce risque.

Un exemple : un double jeu de données

- ▶ $X = (10, 13, 5.5, 11.5, 18, 11.5, 6, 10.5, 6, 6.5, 10, 11.5)$
- ▶ $Y = (1.5, 2.5, 9.5, 6, 6, 4, 4.5)$
- ▶ Un exemple de question qu'on peut se poser : “est-il plausible que ces deux séries de nombres aient été produites selon la même loi de probabilités” ?

Un exemple : un double jeu de données

- ▶ $X = (10, 13, 5.5, 11.5, 18, 11.5, 6, 10.5, 6, 6.5, 10, 11.5)$
- ▶ $Y = (1.5, 2.5, 9.5, 6, 6, 4, 4.5)$
- ▶ Un exemple de question qu'on peut se poser : “est-il plausible que ces deux séries de nombres aient été produites selon la même loi de probabilités” ?
- ▶ On peut calculer une moyenne empirique (10 pour X ; 4.86 pour Y)

Un exemple : un double jeu de données

- ▶ $X = (10, 13, 5.5, 11.5, 18, 11.5, 6, 10.5, 6, 6.5, 10, 11.5)$
- ▶ $Y = (1.5, 2.5, 9.5, 6, 6, 4, 4.5)$
- ▶ Un exemple de question qu'on peut se poser : “est-il plausible que ces deux séries de nombres aient été produites selon la même loi de probabilités” ?
- ▶ On peut calculer une moyenne empirique (10 pour X ; 4.86 pour Y)
- ▶ On peut calculer un écart-type empirique (3.6 pour X ; 2.6 pour Y)

Un exemple : un double jeu de données

- ▶ $X = (10, 13, 5.5, 11.5, 18, 11.5, 6, 10.5, 6, 6.5, 10, 11.5)$
- ▶ $Y = (1.5, 2.5, 9.5, 6, 6, 4, 4.5)$
- ▶ Un exemple de question qu'on peut se poser : “est-il plausible que ces deux séries de nombres aient été produites selon la même loi de probabilités” ?
- ▶ On peut calculer une moyenne empirique (10 pour X ; 4.86 pour Y)
- ▶ On peut calculer un écart-type empirique (3.6 pour X ; 2.6 pour Y)
- ▶ On peut calculer une médiane (10.25 pour X ; 4.5 pour Y)

Un exemple : un double jeu de données

- ▶ $X = (10, 13, 5.5, 11.5, 18, 11.5, 6, 10.5, 6, 6.5, 10, 11.5)$
- ▶ $Y = (1.5, 2.5, 9.5, 6, 6, 4, 4.5)$
- ▶ Un exemple de question qu'on peut se poser : “est-il plausible que ces deux séries de nombres aient été produites selon la même loi de probabilités” ?
- ▶ On peut calculer une moyenne empirique (10 pour X ; 4.86 pour Y)
- ▶ On peut calculer un écart-type empirique (3.6 pour X ; 2.6 pour Y)
- ▶ On peut calculer une médiane (10.25 pour X ; 4.5 pour Y)
- ▶ Peut-on considérer que les différences sont assez importantes pour rejeter l'idée que les deux séries suivent la même loi ?

Un exemple : un double jeu de données

- ▶ $X = (10, 13, 5.5, 11.5, 18, 11.5, 6, 10.5, 6, 6.5, 10, 11.5)$
- ▶ $Y = (1.5, 2.5, 9.5, 6, 6, 4, 4.5)$
- ▶ Un exemple de question qu'on peut se poser : “est-il plausible que ces deux séries de nombres aient été produites selon la même loi de probabilités” ?
- ▶ On peut calculer une moyenne empirique (10 pour X ; 4.86 pour Y)
- ▶ On peut calculer un écart-type empirique (3.6 pour X ; 2.6 pour Y)
- ▶ On peut calculer une médiane (10.25 pour X ; 4.5 pour Y)
- ▶ Peut-on considérer que les différences sont assez importantes pour rejeter l'idée que les deux séries suivent la même loi ?
- ▶ (Après tout, si on tire 5 fois avec un dé équilibré, il est aussi probable d'observer 6, 1, 3, 1, 1 que 6, 6, 6, 6, 6)

Le cadre statistique général

- ▶ On dispose (ou on est capable d'obtenir) d'un *échantillon* : n tirages $X_1, X_2, \dots, X_n$ **indépendants** d'une **même loi** (i.i.d.)

Le cadre statistique général

- ▶ On dispose (ou on est capable d'obtenir) d'un *échantillon* : n tirages $X_1, X_2, \dots, X_n$ **indépendants** d'une **même loi** (i.i.d.)
- ▶ ...mais la loi elle-même, ou certains *paramètres*, sont inconnus

Le cadre statistique général

- ▶ On dispose (ou on est capable d'obtenir) d'un *échantillon* : n tirages $X_1, X_2, \dots, X_n$ **indépendants** d'une **même loi** (i.i.d.)
- ▶ ...mais la loi elle-même, ou certains *paramètres*, sont inconnus
- ▶ ...et on voudrait “dire quelque chose” sur la loi, typiquement répondre à une question :
 - ▶ quelle est **l'espérance** de la loi ?
 - ▶ est-ce que l'espérance vaut au moins 1 ?
 - ▶ est-ce que la loi derrière l'échantillon $X_1, \dots, X_n$ est la même que la loi derrière l'échantillon $Y_1, \dots, Y_m$ (cas avec deux échantillons) ?
 - ▶ (cas où les tirages sont des couples (X_i, Y_i) , donc on a une loi des couples) est-ce que la loi du couple est une loi d'un couple de variables indépendantes ?

Le cadre statistique général

- ▶ On dispose (ou on est capable d'obtenir) d'un *échantillon* : n tirages $X_1, X_2, \dots, X_n$ **indépendants** d'une **même loi** (i.i.d.)
- ▶ ...mais la loi elle-même, ou certains *paramètres*, sont inconnus
- ▶ ...et on voudrait “dire quelque chose” sur la loi, typiquement répondre à une question :
 - ▶ quelle est **l'espérance** de la loi ?
 - ▶ est-ce que l'espérance vaut au moins 1 ?
 - ▶ est-ce que la loi derrière l'échantillon $X_1, \dots, X_n$ est la même que la loi derrière l'échantillon $Y_1, \dots, Y_m$ (cas avec deux échantillons) ?
 - ▶ (cas où les tirages sont des couples (X_i, Y_i) , donc on a une loi des couples) est-ce que la loi du couple est une loi d'un couple de variables indépendantes ?
- ▶ Parfois, on sait que la loi inconnue fait partie d'une famille de lois, mais avec un ou plusieurs paramètres inconnus (ex : loi binomiale ; loi gaussienne ; etc)

Le problème de l'estimation

- La question qu'on se pose : **proposer une valeur crédible** pour un paramètre de la loi.

Le problème de l'estimation

- ▶ La question qu'on se pose : **proposer une valeur crédible** pour un paramètre de la loi.
- ▶ **Exemple** : le paramètre d'une loi géométrique ; l'espérance, en général ; l'écart-type. . .

Le problème de l'estimation

- ▶ La question qu'on se pose : **proposer une valeur crédible** pour un paramètre de la loi.
- ▶ **Exemple** : le paramètre d'une loi géométrique ; l'espérance, en général ; l'écart-type. . .
- ▶ **Démarche** : On se donne une **façon systématique** de fournir une valeur à proposer (pour le paramètre), en fonction de la liste des valeurs de l'échantillon.

Le problème de l'estimation

- ▶ La question qu'on se pose : **proposer une valeur crédible** pour un paramètre de la loi.
- ▶ **Exemple** : le paramètre d'une loi géométrique ; l'espérance, en général ; l'écart-type. . .
- ▶ **Démarche** : On se donne une **façon systématique** de fournir une valeur à proposer (pour le paramètre), en fonction de la liste des valeurs de l'échantillon.
- ▶ On appelle **estimateur**, n'importe quelle fonction (formule) T , dépendant de n et des n valeurs $X_1, \dots, X_n$ de l'échantillon ;

Le problème de l'estimation

- ▶ La question qu'on se pose : **proposer une valeur crédible** pour un paramètre de la loi.
- ▶ **Exemple** : le paramètre d'une loi géométrique ; l'espérance, en général ; l'écart-type. . .
- ▶ **Démarche** : On se donne une **façon systématique** de fournir une valeur à proposer (pour le paramètre), en fonction de la liste des valeurs de l'échantillon.
- ▶ On appelle **estimateur**, n'importe quelle fonction (formule) T , dépendant de n et des n valeurs $X_1, \dots, X_n$ de l'échantillon ;
- ▶ Une fois qu'on a les valeurs $x_1, \dots, x_n$, la valeur $T_n(x_1, \dots, x_n)$ porte le nom d'**estimation**.

Le problème de l'estimation

- ▶ La question qu'on se pose : **proposer une valeur crédible** pour un paramètre de la loi.
- ▶ **Exemple** : le paramètre d'une loi géométrique ; l'espérance, en général ; l'écart-type. . .
- ▶ **Démarche** : On se donne une **façon systématique** de fournir une valeur à proposer (pour le paramètre), en fonction de la liste des valeurs de l'échantillon.
- ▶ On appelle **estimateur**, n'importe quelle fonction (formule) T , dépendant de n et des n valeurs $X_1, \dots, X_n$ de l'échantillon ;
- ▶ Une fois qu'on a les valeurs $x_1, \dots, x_n$, la valeur $T_n(x_1, \dots, x_n)$ porte le nom d'**estimation**.
- ▶ Tel quel, n'importe quoi peut se prétendre estimateur ; simplement, il y en a qui sont meilleurs que d'autres. . . et notre objectif est d'en choisir de bons.

Biais d'un estimateur

- Puisque $X_1, \dots, X_n$ sont aléatoires, $T_n(X_1, \dots, X_n)$ est aussi une variable aléatoire (sa valeur dépend des valeurs prises par les X_i)

Biais d'un estimateur

- ▶ Puisque $X_1, \dots, X_n$ sont aléatoires, $T_n(X_1, \dots, X_n)$ est aussi une variable aléatoire (sa valeur dépend des valeurs prises par les X_i)
- ▶ On peut donc s'intéresser à l'espérance de $T_n(X_1, \dots, X_n)$

Biais d'un estimateur

- ▶ Puisque $X_1, \dots, X_n$ sont aléatoires, $T_n(X_1, \dots, X_n)$ est aussi une variable aléatoire (sa valeur dépend des valeurs prises par les X_i)
- ▶ On peut donc s'intéresser à l'espérance de $T_n(X_1, \dots, X_n)$
- ▶ Si on cherche à estimer un paramètre θ , on dit que T_n **est un estimateur sans biais pour θ si** $\mathbb{E}(T_n(X_1, \dots, X_n)) = \theta$

Biais d'un estimateur

- ▶ Puisque $X_1, \dots, X_n$ sont aléatoires, $T_n(X_1, \dots, X_n)$ est aussi une variable aléatoire (sa valeur dépend des valeurs prises par les X_i)
- ▶ On peut donc s'intéresser à l'espérance de $T_n(X_1, \dots, X_n)$
- ▶ Si on cherche à estimer un paramètre θ , on dit que T_n **est un estimateur sans biais pour θ si** $\mathbb{E}(T_n(X_1, \dots, X_n)) = \theta$
- ▶ (Sinon, $\mathbb{E}(T_n(X_1, \dots, X_n)) - \theta$ est appelé le biais de l'estimateur)

Biais d'un estimateur

- ▶ Puisque $X_1, \dots, X_n$ sont aléatoires, $T_n(X_1, \dots, X_n)$ est aussi une variable aléatoire (sa valeur dépend des valeurs prises par les X_i)
- ▶ On peut donc s'intéresser à l'espérance de $T_n(X_1, \dots, X_n)$
- ▶ Si on cherche à estimer un paramètre θ , on dit que T_n **est un estimateur sans biais pour θ si** $\mathbb{E}(T_n(X_1, \dots, X_n)) = \theta$
- ▶ (Sinon, $\mathbb{E}(T_n(X_1, \dots, X_n)) - \theta$ est appelé le biais de l'estimateur)
- ▶ Un estimateur sans biais est souhaitable, même si ce n'est pas la panacée...

Un exemple d'estimateur

- ▶ **Situation** : on a droit à un échantillon d'une loi, mais on ne sait presque rien dessus ; seulement qu'elle a une espérance
- ▶ On cherche à estimer cette espérance, qu'on note θ

Un exemple d'estimateur

- ▶ **Situation** : on a droit à un échantillon d'une loi, mais on ne sait presque rien dessus ; seulement qu'elle a une espérance
- ▶ On cherche à estimer cette espérance, qu'on note θ
- ▶ X_1 est un estimateur sans biais (!)

Un exemple d'estimateur

- ▶ **Situation** : on a droit à un échantillon d'une loi, mais on ne sait presque rien dessus ; seulement qu'elle a une espérance
- ▶ On cherche à estimer cette espérance, qu'on note θ
- ▶ X_1 est un estimateur sans biais (!)
- ▶ $(X_1 + X_2)/2$ est aussi un estimateur sans biais

Un exemple d'estimateur

- ▶ **Situation** : on a droit à un échantillon d'une loi, mais on ne sait presque rien dessus ; seulement qu'elle a une espérance
- ▶ On cherche à estimer cette espérance, qu'on note θ
- ▶ X_1 est un estimateur sans biais (!)
- ▶ $(X_1 + X_2)/2$ est aussi un estimateur sans biais
- ▶ $(X_1 + 2X_2 + 3X_3 + 4X_4)/10$ est aussi un estimateur sans biais

Un exemple d'estimateur

- ▶ **Situation** : on a droit à un échantillon d'une loi, mais on ne sait presque rien dessus ; seulement qu'elle a une espérance
- ▶ On cherche à estimer cette espérance, qu'on note θ
- ▶ X_1 est un estimateur sans biais (!)
- ▶ $(X_1 + X_2)/2$ est aussi un estimateur sans biais
- ▶ $(X_1 + 2X_2 + 3X_3 + 4X_4)/10$ est aussi un estimateur sans biais
- ▶ $(X_1 + 1)/2$ est un estimateur biaisé (sauf si $\theta = 1$)

Un exemple d'estimateur

- ▶ **Situation** : on a droit à un échantillon d'une loi, mais on ne sait presque rien dessus ; seulement qu'elle a une espérance
- ▶ On cherche à estimer cette espérance, qu'on note θ
- ▶ X_1 est un estimateur sans biais (!)
- ▶ $(X_1 + X_2)/2$ est aussi un estimateur sans biais
- ▶ $(X_1 + 2X_2 + 3X_3 + 4X_4)/10$ est aussi un estimateur sans biais
- ▶ $(X_1 + 1)/2$ est un estimateur biaisé (sauf si $\theta = 1$)
- ▶ Un estimateur “naturel” pour l'espérance, et qui utilise tout l'échantillon : la “moyenne empirique” $(X_1 + \dots + X_n)/n$.

Estimateur convergent

- ▶ Deuxième propriété souhaitable d'un estimateur

Estimateur convergent

- ▶ Deuxième propriété souhaitable d'un estimateur
- ▶ En fait on considère une suite d'estimateurs T_n :
potentiellement, un pour chaque taille d'échantillon

Estimateur convergent

- ▶ Deuxième propriété souhaitable d'un estimateur
- ▶ En fait on considère une suite d'estimateurs T_n :
potentiellement, un pour chaque taille d'échantillon
- ▶ Un estimateur $(T_n)_{n \geq 1}$ pour un paramètre θ , est **convergent** si T_n converge en probabilités vers θ

Estimateur convergent

- ▶ Deuxième propriété souhaitable d'un estimateur
- ▶ En fait on considère une suite d'estimateurs T_n :
potentiellement, un pour chaque taille d'échantillon
- ▶ Un estimateur $(T_n)_{n \geq 1}$ pour un paramètre θ , est **convergent** si T_n converge en probabilités vers θ
- ▶ **Exemple symptomatique** : $T_n = (X_1 + X_2 + \cdots + X_n)/n$ est un estimateur convergent pour l'espérance (par la loi des grands nombres)

Estimateur convergent

- ▶ Deuxième propriété souhaitable d'un estimateur
- ▶ En fait on considère une suite d'estimateurs T_n : potentiellement, un pour chaque taille d'échantillon
- ▶ Un estimateur $(T_n)_{n \geq 1}$ pour un paramètre θ , est **convergent** si T_n converge en probabilités vers θ
- ▶ **Exemple symptomatique** : $T_n = (X_1 + X_2 + \dots + X_n)/n$ est un estimateur convergent pour l'espérance (par la loi des grands nombres)
- ▶ Un estimateur convergent est une bonne chose : par définition, "si on prend un échantillon assez grand, la probabilité que l'estimation soit loin du vrai paramètre, finira par passer en-dessous de n'importe quel seuil"

Estimateur convergent

- ▶ Deuxième propriété souhaitable d'un estimateur
- ▶ En fait on considère une suite d'estimateurs T_n : potentiellement, un pour chaque taille d'échantillon
- ▶ Un estimateur $(T_n)_{n \geq 1}$ pour un paramètre θ , est **convergent** si T_n converge en probabilités vers θ
- ▶ **Exemple symptomatique** : $T_n = (X_1 + X_2 + \dots + X_n)/n$ est un estimateur convergent pour l'espérance (par la loi des grands nombres)
- ▶ Un estimateur convergent est une bonne chose : par définition, "si on prend un échantillon assez grand, la probabilité que l'estimation soit loin du vrai paramètre, finira par passer en-dessous de n'importe quel seuil"
- ▶ (Par contre, la convergence en probabilités ne fournit aucune garantie de "quelle taille il faut prendre" : pour ça, il faut un contrôle sur la "vitesse de convergence")

Estimation de variance

- ▶ Même hypothèse que précédemment : la loi des X_i admet une espérance m , et une variance σ^2 ; on voudrait estimer σ^2 .

Estimation de variance

- ▶ Même hypothèse que précédemment : la loi des X_i admet une espérance m , et une variance σ^2 ; on voudrait estimer σ^2 .
- ▶ Comme on a naturellement $\bar{X}_n = (X_1 + \cdots + X_n)/n$ qui est un estimateur sans biais de m , on pense à remplacer m par $\bar{X}_n$ dans la formule de la variance, et à prendre

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

Estimation de variance

- ▶ Même hypothèse que précédemment : la loi des X_i admet une espérance m , et une variance σ^2 ; on voudrait estimer σ^2 .
- ▶ Comme on a naturellement $\bar{X}_n = (X_1 + \cdots + X_n)/n$ qui est un estimateur sans biais de m , on pense à remplacer m par $\bar{X}_n$ dans la formule de la variance, et à prendre

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

- ▶ Cet estimateur (de σ^2) est-il sans biais ? On calcule son espérance. . .

Estimation de variance

- ▶ Même hypothèse que précédemment : la loi des X_i admet une espérance m , et une variance σ^2 ; on voudrait estimer σ^2 .
- ▶ Comme on a naturellement $\bar{X}_n = (X_1 + \cdots + X_n)/n$ qui est un estimateur sans biais de m , on pense à remplacer m par $\bar{X}_n$ dans la formule de la variance, et à prendre

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

- ▶ Cet estimateur (de σ^2) est-il sans biais ? On calcule son espérance. . .
- ▶ (Calcul au tableau)

Estimation de variance

- ▶ Même hypothèse que précédemment : la loi des X_i admet une espérance m , et une variance σ^2 ; on voudrait estimer σ^2 .
- ▶ Comme on a naturellement $\bar{X}_n = (X_1 + \dots + X_n)/n$ qui est un estimateur sans biais de m , on pense à remplacer m par $\bar{X}_n$ dans la formule de la variance, et à prendre

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

- ▶ Cet estimateur (de σ^2) est-il sans biais ? On calcule son espérance. . .
- ▶ (Calcul au tableau)
- ▶ On s'aperçoit que $\mathbb{E}(S_n) = \frac{n-1}{n}\sigma^2$: **ce n'est pas un estimateur sans biais** (l'espérance n'est pas égale à σ^2)

Estimation de variance

- ▶ Même hypothèse que précédemment : la loi des X_i admet une espérance m , et une variance σ^2 ; on voudrait estimer σ^2 .
- ▶ Comme on a naturellement $\bar{X}_n = (X_1 + \cdots + X_n)/n$ qui est un estimateur sans biais de m , on pense à remplacer m par $\bar{X}_n$ dans la formule de la variance, et à prendre

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

- ▶ Cet estimateur (de σ^2) est-il sans biais ? On calcule son espérance. . .
- ▶ (Calcul au tableau)
- ▶ On s'aperçoit que $\mathbb{E}(S_n) = \frac{n-1}{n}\sigma^2$: **ce n'est pas un estimateur sans biais** (l'espérance n'est pas égale à σ^2)
- ▶ **Mais** si on multiplie par $\frac{n}{n-1}$, ça devient sans biais !

Estimation de variance

- ▶ Même hypothèse que précédemment : la loi des X_i admet une espérance m , et une variance σ^2 ; on voudrait estimer σ^2 .
- ▶ Comme on a naturellement $\bar{X}_n = (X_1 + \dots + X_n)/n$ qui est un estimateur sans biais de m , on pense à remplacer m par $\bar{X}_n$ dans la formule de la variance, et à prendre

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

- ▶ Cet estimateur (de σ^2) est-il sans biais ? On calcule son espérance. . .
- ▶ (Calcul au tableau)
- ▶ On s'aperçoit que $\mathbb{E}(S_n) = \frac{n-1}{n}\sigma^2$: **ce n'est pas un estimateur sans biais** (l'espérance n'est pas égale à σ^2)
- ▶ **Mais** si on multiplie par $\frac{n}{n-1}$, ça devient sans biais !
- ▶ **Conclusion** : on a un estimateur sans biais de la variance, sous la forme de

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

Notion d'intervalle de confiance

- ▶ Notion dont la **signification** est souvent mal comprise.

Notion d'intervalle de confiance

- ▶ Notion dont la **signification** est souvent mal comprise.
- ▶ Au lieu de chercher seulement à fournir une estimation d'un paramètre, on inclut une marge d'erreur : on donne un **intervalle** (aux extrêmités définies par des variables aléatoires) dont les extrêmités sont probablement de part et d'autres du paramètre à estimer.

Notion d'intervalle de confiance

- ▶ Notion dont la **signification** est souvent mal comprise.
- ▶ Au lieu de chercher seulement à fournir une estimation d'un paramètre, on inclut une marge d'erreur : on donne un **intervalle** (aux extrêmités définies par des variables aléatoires) dont les extrêmités sont probablement de part et d'autres du paramètre à estimer.
- ▶ **Remarque** : la condition $|T_n - \theta| \leq a$ peut se réécrire de manière équivalente de deux manières :

Notion d'intervalle de confiance

- ▶ Notion dont la **signification** est souvent mal comprise.
- ▶ Au lieu de chercher seulement à fournir une estimation d'un paramètre, on inclut une marge d'erreur : on donne un **intervalle** (aux extrêmités définies par des variables aléatoires) dont les extrêmités sont probablement de part et d'autres du paramètre à estimer.
- ▶ **Remarque** : la condition $|T_n - \theta| \leq a$ peut se réécrire de manière équivalente de deux manières :
 - ▶ $T_n \in [\theta - a, \theta + a]$

Notion d'intervalle de confiance

- ▶ Notion dont la **signification** est souvent mal comprise.
- ▶ Au lieu de chercher seulement à fournir une estimation d'un paramètre, on inclut une marge d'erreur : on donne un **intervalle** (aux extrêmités définies par des variables aléatoires) dont les extrêmités sont probablement de part et d'autres du paramètre à estimer.
- ▶ **Remarque** : la condition $|T_n - \theta| \leq a$ peut se réécrire de manière équivalente de deux manières :
 - ▶ $T_n \in [\theta - a, \theta + a]$
 - ▶ $\theta \in [T_n - a, T_n + a]$

Notion d'intervalle de confiance

- ▶ Notion dont la **signification** est souvent mal comprise.
- ▶ Au lieu de chercher seulement à fournir une estimation d'un paramètre, on inclut une marge d'erreur : on donne un **intervalle** (aux extrêmités définies par des variables aléatoires) dont les extrêmités sont probablement de part et d'autres du paramètre à estimer.
- ▶ **Remarque** : la condition $|T_n - \theta| \leq a$ peut se réécrire de manière équivalente de deux manières :
 - ▶ $T_n \in [\theta - a, \theta + a]$
 - ▶ $\theta \in [T_n - a, T_n + a]$
- ▶ **Donc**, si on peut prouver que l'estimateur T_n est probablement proche de θ , on prouve que θ est probablement dans l'intervalle $[T_n - a, T_n + a]$

Notion d'intervalle de confiance

- ▶ Notion dont la **signification** est souvent mal comprise.
- ▶ Au lieu de chercher seulement à fournir une estimation d'un paramètre, on inclut une marge d'erreur : on donne un **intervalle** (aux extrêmités définies par des variables aléatoires) dont les extrêmités sont probablement de part et d'autres du paramètre à estimer.
- ▶ **Remarque** : la condition $|T_n - \theta| \leq a$ peut se réécrire de manière équivalente de deux manières :
 - ▶ $T_n \in [\theta - a, \theta + a]$
 - ▶ $\theta \in [T_n - a, T_n + a]$
- ▶ **Donc**, si on peut prouver que l'estimateur T_n est probablement proche de θ , on prouve que θ est probablement dans l'intervalle $[T_n - a, T_n + a]$
- ▶ **Définition** : pour $0 < \alpha < 1$, on appelle *intervalle de confiance au niveau de risque α* pour le paramètre θ , la donnée de deux fonctions $A_n(X_1, \dots, X_n)$ et $B_n(X_1, \dots, X_n)$ telle qu'on ait

$$\mathbb{P}(\theta \in [A_n, B_n]) \geq 1 - \alpha$$

Un exemple : paramètre d'une Bernoulli

- ▶ On cherche à estimer le paramètre p d'une loi de Bernoulli

Un exemple : paramètre d'une Bernoulli

- ▶ On cherche à estimer le paramètre p d'une loi de Bernoulli
- ▶ On suppose donc qu'on a un échantillon $B_1, \dots, B_n$

Un exemple : paramètre d'une Bernoulli

- ▶ On cherche à estimer le paramètre p d'une loi de Bernoulli
- ▶ On suppose donc qu'on a un échantillon $B_1, \dots, B_n$
- ▶ On prend comme estimateur de p , $\bar{B}_n = (B_1 + \dots + B_n)/n$

Un exemple : paramètre d'une Bernoulli

- ▶ On cherche à estimer le paramètre p d'une loi de Bernoulli
- ▶ On suppose donc qu'on a un échantillon $B_1, \dots, B_n$
- ▶ On prend comme estimateur de p , $\bar{B}_n = (B_1 + \dots + B_n)/n$
- ▶ $n\bar{B}_n$ est binomiale (paramètres n et p , p inconnu), donc d'espérance np et de variance $np(1 - p)$

Un exemple : paramètre d'une Bernoulli

- ▶ On cherche à estimer le paramètre p d'une loi de Bernoulli
- ▶ On suppose donc qu'on a un échantillon $B_1, \dots, B_n$
- ▶ On prend comme estimateur de p , $\bar{B}_n = (B_1 + \dots + B_n)/n$
- ▶ $n\bar{B}_n$ est binomiale (paramètres n et p , p inconnu), donc d'espérance np et de variance $np(1 - p)$
- ▶ donc $\bar{B}_n$ est un estimateur sans biais, de variance $p(1 - p)/n \leq 1/(4n)$

Un exemple : paramètre d'une Bernoulli

- ▶ On cherche à estimer le paramètre p d'une loi de Bernoulli
- ▶ On suppose donc qu'on a un échantillon $B_1, \dots, B_n$
- ▶ On prend comme estimateur de p , $\bar{B}_n = (B_1 + \dots + B_n)/n$
- ▶ $n\bar{B}_n$ est binomiale (paramètres n et p , p inconnu), donc d'espérance np et de variance $np(1 - p)$
- ▶ donc $\bar{B}_n$ est un estimateur sans biais, de variance $p(1 - p)/n \leq 1/(4n)$
- ▶ donc (sans connaître p) on peut affirmer que l'écart-type de $\bar{B}_n$ est au plus égal à $1/2\sqrt{n}$

Un exemple : paramètre d'une Bernoulli

- ▶ On cherche à estimer le paramètre p d'une loi de Bernoulli
- ▶ On suppose donc qu'on a un échantillon $B_1, \dots, B_n$
- ▶ On prend comme estimateur de p , $\bar{B}_n = (B_1 + \dots + B_n)/n$
- ▶ $n\bar{B}_n$ est binomiale (paramètres n et p , p inconnu), donc d'espérance np et de variance $np(1 - p)$
- ▶ donc $\bar{B}_n$ est un estimateur sans biais, de variance $p(1 - p)/n \leq 1/(4n)$
- ▶ donc (sans connaître p) on peut affirmer que l'écart-type de $\bar{B}_n$ est au plus égal à $1/2\sqrt{n}$
- ▶ et pour un risque α , on peut prendre $a = \frac{1}{2\sqrt{\alpha n}}$ (Tchebycheff)

Un exemple : paramètre d'une Bernoulli

- ▶ On cherche à estimer le paramètre p d'une loi de Bernoulli
- ▶ On suppose donc qu'on a un échantillon $B_1, \dots, B_n$
- ▶ On prend comme estimateur de p , $\bar{B}_n = (B_1 + \dots + B_n)/n$
- ▶ $n\bar{B}_n$ est binomiale (paramètres n et p , p inconnu), donc d'espérance np et de variance $np(1-p)$
- ▶ donc $\bar{B}_n$ est un estimateur sans biais, de variance $p(1-p)/n \leq 1/(4n)$
- ▶ donc (sans connaître p) on peut affirmer que l'écart-type de $\bar{B}_n$ est au plus égal à $1/(2\sqrt{n})$
- ▶ et pour un risque α , on peut prendre $a = \frac{1}{2\sqrt{\alpha n}}$ (Tchebycheff)
- ▶ **Résultat** : on a un intervalle de confiance au niveau de risque α avec $[\bar{B}_n - \frac{1}{2\sqrt{\alpha n}}, \bar{B}_n + \frac{1}{2\sqrt{\alpha n}}]$.

Exemple pratique

- ▶ Échantillon de taille 1000 : on tire 1000 Bernoulli, et on observe que 325 donnent la valeur 1 ; on veut un intervalle de confiance au niveau de risque 5% (valeur classique)
- ▶ On a donc $\bar{b}_n = 0.325$
- ▶ En appliquant l'intervalle de confiance basé sur Tchebycheff : on prend un intervalle $[\bar{b}_n - a, \bar{b}_n + a]$, en s'arrangeant pour que l'on ait $\frac{1/4}{1000a^2} = \alpha$
- ▶ On résoud : ça donne $a \simeq 0.07$.
- ▶ L'intervalle de confiance est donc $[0.255, 0.395]$ (une marge d'erreur de ± 0.07 , c'est assez élevé ; et pour la diviser par 2, il faut multiplier par 4 la taille de l'échantillon).

Remarque sur l'interprétation

- ▶ On applique un intervalle de confiance, et on obtient (en appliquant les formules qu'on a définies, avec les valeurs de l'échantillon) une estimation $[a_n, b_n]$

Remarque sur l'interprétation

- ▶ On applique un intervalle de confiance, et on obtient (en appliquant les formules qu'on a définies, avec les valeurs de l'échantillon) une estimation $[a_n, b_n]$
- ▶ **Il n'est pas correct** de dire “la probabilité que le paramètre soit entre a_n et b_n est d'au moins $1 - \alpha$ ”

Remarque sur l'interprétation

- ▶ On applique un intervalle de confiance, et on obtient (en appliquant les formules qu'on a définies, avec les valeurs de l'échantillon) une estimation $[a_n, b_n]$
- ▶ **Il n'est pas correct** de dire “la probabilité que le paramètre soit entre a_n et b_n est d'au moins $1 - \alpha$ ”
- ▶ Une fois qu'on a fait les calculs, a_n et b_n ne sont plus aléatoires : ils valent certaines valeurs, et soit θ est entre les deux, soit il ne l'est pas ; ça n'a plus de probabilité.

Remarque sur l'interprétation

- ▶ On applique un intervalle de confiance, et on obtient (en appliquant les formules qu'on a définies, avec les valeurs de l'échantillon) une estimation $[a_n, b_n]$
- ▶ **Il n'est pas correct** de dire “la probabilité que le paramètre soit entre a_n et b_n est d'au moins $1 - \alpha$ ”
- ▶ Une fois qu'on a fait les calculs, a_n et b_n ne sont plus aléatoires : ils valent certaines valeurs, et soit θ est entre les deux, soit il ne l'est pas ; ça n'a plus de probabilité.
- ▶ L'interprétation correcte : “J'ai appliqué une procédure qui a probabilité au moins $1 - \alpha$ de donner un intervalle contenant θ , et l'intervalle obtenu était $[a_n, b_n]$ ”

Remarque sur l'interprétation

- ▶ On applique un intervalle de confiance, et on obtient (en appliquant les formules qu'on a définies, avec les valeurs de l'échantillon) une estimation $[a_n, b_n]$
- ▶ **Il n'est pas correct** de dire “la probabilité que le paramètre soit entre a_n et b_n est d'au moins $1 - \alpha$ ”
- ▶ Une fois qu'on a fait les calculs, a_n et b_n ne sont plus aléatoires : ils valent certaines valeurs, et soit θ est entre les deux, soit il ne l'est pas ; ça n'a plus de probabilité.
- ▶ L'interprétation correcte : “J'ai appliqué une procédure qui a probabilité au moins $1 - \alpha$ de donner un intervalle contenant θ , et l'intervalle obtenu était $[a_n, b_n]$ ”
- ▶ (Si on réalise un grand nombre de sondages qui fournissent tous un intervalle de confiance au risque 5%, il faut s'attendre à ce qu'un sur 20 se trompe)