

Chapitre 1  
Introduction à la statistique avec R

Définitions

1

Pr. Bruno Falissard

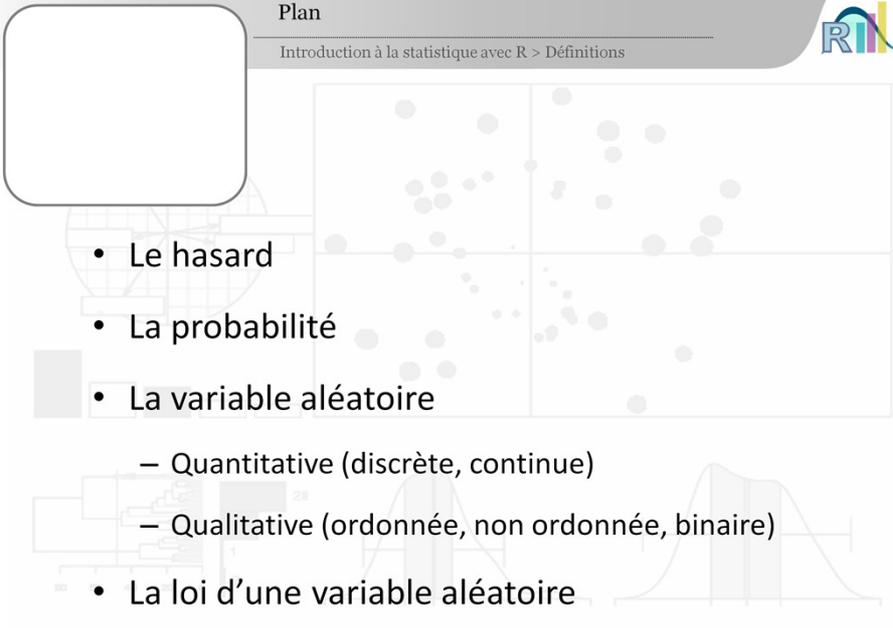
UNIVERSITÉ PARIS SUD

CC BY NC SA

[0:01] Comme la plupart des disciplines, les statistiques ont leur jargon, c'est-à-dire des éléments de vocabulaire qu'il faut connaître pour pouvoir s'y retrouver.

Plan

Introduction à la statistique avec R > Définitions



- Le hasard
- La probabilité
- La variable aléatoire
  - Quantitative (discrète, continue)
  - Qualitative (ordonnée, non ordonnée, binaire)
- La loi d'une variable aléatoire

CC BY NC SA

2

Pr. Bruno Falissard

UNIVERSITÉ PARIS SUD

[0:08] Nous allons voir rapidement la notion

- de hasard,
- de probabilité,
- puis celle de variable aléatoire, assez technique, avec les variables aléatoires quantitatives et celles qui sont qualitatives.
- Nous verrons enfin la notion de loi d'une variable aléatoire en nous focalisant sur les variables aléatoires suivant une loi normale.

Le hasard

Introduction à la statistique avec R > Définitions

- **Le hasard :**
  - Il est la traduction de notre ignorance...
  - Le hasard est donc relatif

3

Pr. Bruno Falissard

UNIVERSITE PARIS SUD

[0:28] Définir le **hasard**, c'est plus une question de philosophie qu'une question de mathématiques. Et au demeurant, ce n'est pas si simple que ça. Pour être très pratique, on dira que le hasard, c'est la traduction de notre ignorance. Quand vous jouez à pile ou face d'ailleurs, c'est bien parce que vous ne savez pas à l'avance de quel côté va tomber la pièce qu'on dit que c'est un jeu de hasard et du coup, ça a comme corollaire que le hasard est une notion relative. Imaginez qu'un jour, en filmant les pièces et en ayant des calculs physiques très sophistiqués, on arrive à prévoir à l'avance où va tomber la pièce, alors on ne pourra plus considérer que le jeu de pile ou face est un jeu de hasard.

La probabilité

Introduction à la statistique avec R > Définitions

- Le hasard
- La probabilité :
  - « Une » ou « des » probabilités ?...
  - Physico-probabilités (fréquence limite)
  - Psycho-probabilités (plausibilité)

4

Pr. Bruno Falissard

UNIVERSITÉ PARIS SUD

CC BY NC SA

[1:06] En statistiques, on utilise tout le temps le mot **probabilité**. On pourrait avoir l'impression que définir une probabilité, c'est simple et c'est mathématique. Et là aussi, ce n'est pas tout à fait le cas. En pratique, il y a deux façons de définir une probabilité.

La première, c'est de considérer que la probabilité, c'est la fréquence d'apparition d'un évènement. Ainsi, la probabilité qu'il pleuve un jour donné à Biarritz, c'est égal au nombre de jours où il a plu, par exemple dans les dix dernières années, sur le nombre de jours qu'il y a eu dans les dix dernières années, c'est-à-dire 3650. Dans ce cas là, certains parlent de physico-probabilité ou de fréquence limite.

Mais il y a des situations où en fait ça n'a pas de sens de parler de probabilité en ces termes. Imaginez que je sois en vacances à Biarritz, et que je me demande le soir quand je prépare mes vêtements pour le lendemain : "Quelle est la probabilité qu'il pleuve demain ?" Mais la probabilité qu'il pleuve demain, demain c'est un jour unique dans l'histoire de l'humanité. Ce n'est donc pas un élément reproductible et on ne peut pas parler de fréquence limite. Dans ces cas-là, on parle plutôt de plausibilité. Certains même utilisent même le terme de psycho-probabilité. En pratique, en statistiques, dans les modèles, on a indifféremment recours à l'une ou l'autre de ces définitions de la probabilité.

La variable aléatoire

Introduction à la statistique avec R > Définitions

- Le hasard
- La probabilité
- **La variable aléatoire**
  - Quantitative (discrète, continue)
  - Qualitative (ordonnée, non ordonnée, binaire)

5

Pr. Bruno Falissard

UNIVERSITÉ PARIS SUD

[2:27] Les statisticiens utilisent souvent le mot **variable aléatoire**. Qu'est-ce qu'il signifie ? Une variable, c'est tout simplement quelque chose qu'on a mesuré et elle est aléatoire si le résultat de la mesure est en partie dû au hasard. On oppose souvent les variables aléatoires quantitatives à celles qui sont qualitatives.

Une variable est **quantitative** quand ça a un sens de faire la somme ou la différence de plusieurs résultats. Par exemple, vous prenez le poids ou la taille, ça a un sens de faire une différence de poids ou de taille. Donc, on dit que ces mesures sont des mesures quantitatives. Parmi les mesures quantitatives, il y a celles qui sont discrètes et celles qui sont continues. Elles sont **discrètes** quand il y a un nombre limité de résultats possibles. Elles sont **continues**, au contraire, quand le nombre de résultats possibles est très grand, voire infini. Et on peut avoir des surprises en pratique sur ce qui est quantitatif continu et ce qui est quantitatif discret. Par exemple, si on prend un cas médical, la numération globulaire, c'est-à-dire le nombre de globules rouges par millimètre cube de sang. Vu de loin, on a l'impression que c'est une variable quantitative discrète puisque c'est un comptage. En réalité, le nombre de résultats possibles est tellement grand – un résultat possible peut être 5 000 000, 5 500 000, 5 250 000, etc. il y a quasiment autant de numérations possibles que de sujets – et donc en pratique, pour l'analyse statistique, la numération globulaire est une variable quantitative continue, au contraire de la tension artérielle mesurée au brassard : si vous prenez le nombre le plus bas, il peut valoir 6 7 7,5 8 8,5 9. C'est une variable aléatoire quantitative discrète parce qu'il y a un nombre limité de mesures. Alors que pour un physicien, une pression artérielle, une tension, c'est quelque chose de continu. Mais pour le statisticien, ce qui compte, c'est la réalité de la mesure qui a été faite. Et dans la réalité, la tension artérielle mesurée par le médecin, c'est quantitatif discret.

La variable aléatoire

Introduction à la statistique avec R > Définitions

- Le hasard
- La probabilité
- **La variable aléatoire**
  - Quantitative (discrète, continue)
  - Qualitative (ordonnée, non ordonnée, binaire)

5'

Pr. Bruno Falissard

[4:33] Alors à propos des variables aléatoires **qualitatives**, on ne peut pas en faire la somme et la différence. Pour rester dans le domaine médical, par exemple, le groupe sanguin (A, B, O, AB), c'est une variable aléatoire qualitative.

Il y a des variables aléatoires qualitatives qui sont **ordonnées**. Par exemple, le niveau de satisfaction que l'on a du Président de la République. On peut dire qu'on est "pas satisfait du tout", "un peu satisfait", "moyennement satisfait", "très satisfait". On va coder cela en 0, 1, 2, 3. Ça ne sera pas vraiment une variable aléatoire quantitative parce qu'on ne peut pas faire la somme ou la différence aussi simplement que ça de résultats de satisfaction qui sont codés en "pas du tout", "un peu", "beaucoup", "passionnément". Néanmoins, "un peu" est compris entre "pas du tout" et "moyennement" et donc on n'est pas vraiment dans une variable aléatoire qualitative pure. C'est pour ça qu'on parle de variable aléatoire qualitative ordonnée.

Alors il y a un cas particulier très important de variables aléatoires qualitatives. Ce sont celles qui sont **binaires** : être au chômage (Oui/Non), être à la retraite (Oui/Non), être majeur (Oui/Non). Dans la plupart des disciplines, les variables aléatoires binaires sont très souvent utilisées et les statisticiens ont construit des modèles spécifiques pour les étudier.

La loi d'une variable aléatoire

Introduction à la statistique avec R > Définitions



- Le hasard
- La probabilité
- La variable aléatoire
  - Quantitative (discrète, continue)
  - Qualitative (ordonnée, non ordonnée, binaire)
- **La loi d'une variable aléatoire**



6

Pr. Bruno Falissard



[5:47] La **loi d'une variable aléatoire**, c'est la liste des probabilités de chacune des valeurs qu'elle peut prendre.

La loi d'une variable aléatoire

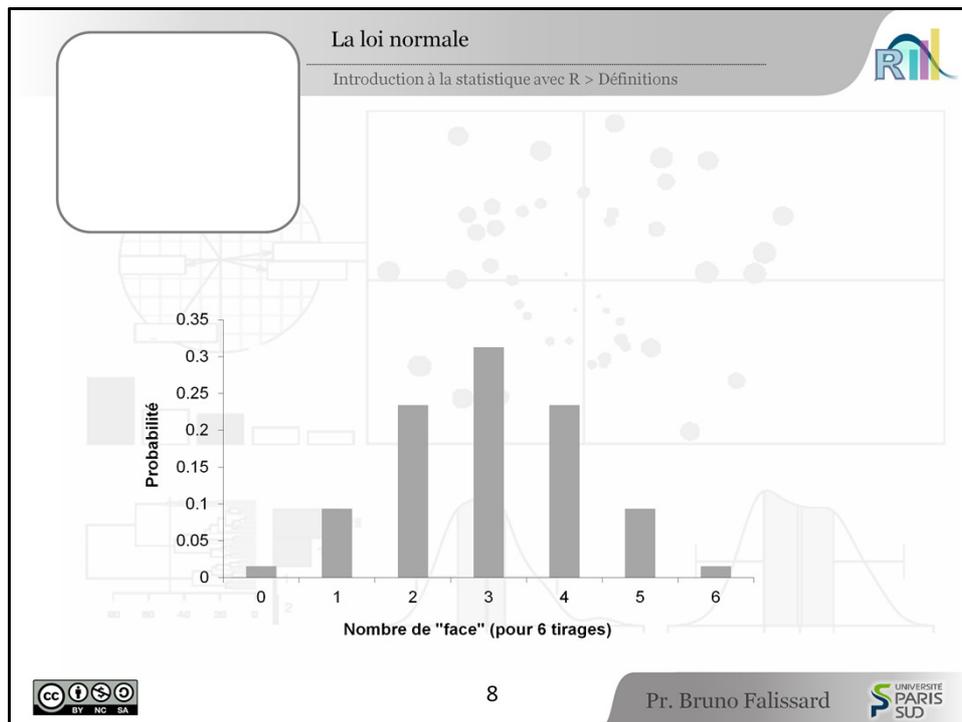
Introduction à la statistique avec R > Définitions

- Exemple du jeu de dé :
  - Probabilité d'obtenir « 1 »  $\rightarrow 1/6$
  - Probabilité d'obtenir « 2 »  $\rightarrow 1/6$
  - ...
  - Probabilité d'obtenir « 6 »  $\rightarrow 1/6$

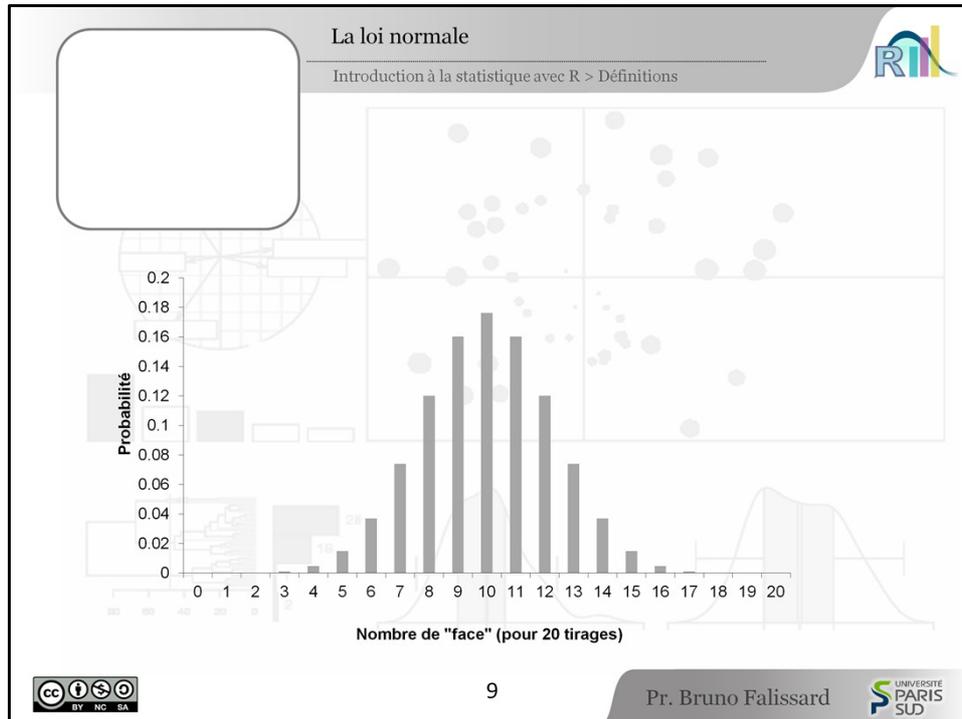
7

Pr. Bruno Falissard

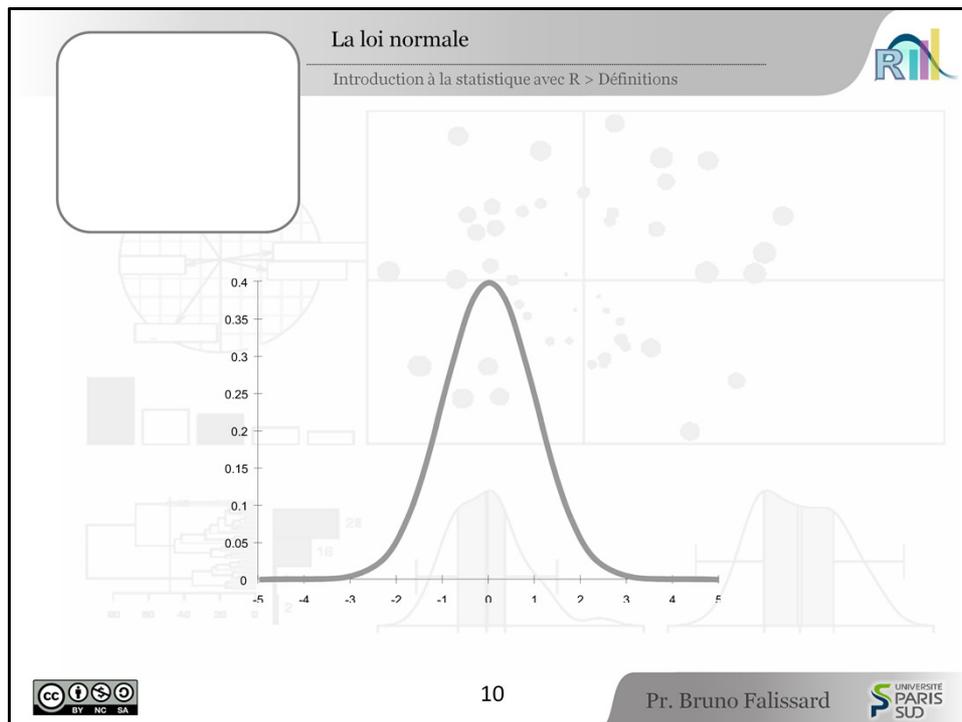
[5:57] Prenons un exemple classique, celui du dé que l'on jette sur une table de Casino. Le résultat possible du jet de dé, c'est 1, 2, 3, 4, 5 ou 6. Et si le dé est un dé qui n'est pas pipé – en termes de statisticien, on dit qu'il est équiprobable – alors la probabilité d'obtenir "1", c'est 1 chance sur 6. Et c'est la même chose pour 2, 3, 4, 5, 6. On dit donc que la loi de la variable aléatoire "jeu du dé à 6 faces" est un  $1/6, 1/6, 1/6, 1/6, 1/6, 1/6$ .



[6:30] Prenons maintenant un nouvel exemple de distribution de variable aléatoire. On va jouer à pile ou face. On va jouer 6 fois. On va compter le nombre total de "face" que l'on fait. Si on a un peu de chance, ça devrait tourner autour de 2, 3 ou 4. Et puis exceptionnellement, ça peut être 0 ou 6. Nous avons sur la figure suivante la distribution de probabilité calculée exactement pour chacune des possibilités.



[6:55] Re commençons maintenant notre expérience et au lieu de jeter 6 fois la pièce, nous allons la jeter 20 fois. On comptabilise toujours le nombre total de "face". Ce nombre peut varier de 0 à 20, en général il sera autour de 10, exceptionnellement 4 et très rarement 0, 1, ou 19 et 20. La distribution de cette nouvelle variable aléatoire est présentée sur la diapositive suivante. Elle prend une forme très régulière et harmonieuse. Et d'ailleurs à la limite, quand le nombre de tirages de pièces, plutôt que d'être égal à 20 tend vers l'infini, la loi tend à se rapprocher d'une courbe continue que l'on dénomme courbe de "Gauss" ou courbe "normale".



[7:36] La loi normale a une grande importance en statistiques. En effet, quand une variable est la résultante d'un grand nombre de variables aléatoires indépendantes, alors cette variable suit une loi normale. Par exemple, la taille d'un individu, elle est la résultante de plusieurs facteurs génétiques, de facteurs environnementaux, et même sociaux et culturels. Tous ces facteurs étant plus ou moins indépendants, la taille d'un individu suit finalement une loi normale. Partant de cette constatation, les statisticiens ont développé des tests optimaux, les plus performants possibles pour des variables suivant des lois normales. Bien entendu, dans la suite du cours, nous ne manquerons pas de les étudier.