3.6 – Estimation sans biais de variance minimale

Borne de Cramer Rao

Soit un échantillon $\{X_1, X_2, \ldots, X_n\}$ dont la loi dépend d'un paramètre θ.

Soit T_n un estimateur **non biaisé** d'une fonction du paramètre $\theta,\ g(\theta)$.

La borne de Cramer Rao est définie par : $BCR(\theta) = \frac{[g\prime(\theta)]^2}{I_n(\theta)}$

En pratique, lorsqu'on s'intéresse à **une estimation de $\boldsymbol{\theta}$**, alors

$$\boxed{BCR(\theta) = \frac{1}{I_n(\theta)}}$$

Inégalité de Fréchet Darmois Cramer Rao (FDCR)

Soit T_n un estimateur **sans biais** de $g(\theta)$

L'inégalité de FDCR fournit une borne inférieure (borne FDCR) sur la variance d'un estimateur sans biais, basée sur l'information de Fisher :

$$V(T_n) \geq BCR(\theta) = \frac{[g'(\theta)]^2}{I_n(\theta)}$$

Lorsque T_n est **un estimateur sans biais de θ**, alors

$$\boxed{V(T_n) \geq BCR(\theta) = \frac{1}{I_n(\theta)}}$$

Efficacité d'un estimateur

Un estimateur **sans biais** T_n d'un paramètre θ est dit **efficace** si sa variance $V(T_n)$ atteint la borne FDCR :

$$\boxed{V(T_n) = BCR(\theta) = \frac{1}{I_n(\theta)}}$$

Remarque

La borne FDCR ne peut être atteinte que si la loi de X est de **forme exponentielle**

Exemples

- Efficacité de l'estimateur $\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$ du paramètre σ^2 d'une loi normale ?
- Efficacité de l'estimateur $\bar{X}$ du paramètre μ d'une loi normale ?

3.7 – Normalité asymptotique

Théorème Central Limite

Soit (X_n) une suite de variables aléatoires de même loi d'espérance μ et de variance σ^2. Alors,

$$\frac{1}{\sqrt{n}}\left(\frac{X_1 + \cdots + X_n - n\mu}{\sigma}\right) \xrightarrow{n\to\infty} \mathcal{N}(0,1)$$

ou encore,

$$\sqrt{n}\,(\bar{X} - \mu) \xrightarrow{n\to\infty} \mathcal{N}(0,\sigma)$$

Normalité asymptotique d'un estimateur

Un estimateur T_n d'un paramètre θ est dit **asymptotiquement normal** si :

$$\sqrt{n}\,(T_n - \theta) \xrightarrow{n\to\infty} \mathcal{N}\big(0, J(\theta)\big) \quad \text{où} \quad J(\theta) > 0$$

Exemples

- Soit un *n* échantillon $\{X_1, X_2, \dots, X_n\}$ de loi de Bernoulli (θ)
  Montrer que $\bar{X}_n$ est asymptotiquement normal et donner $J(\theta)$
- Idem pour une loi exponentielle $\mathcal{E}\left(\frac{1}{\lambda}\right)$

Normalité asymptotique de l'estimateur du MV

Soit un *n* échantillon $\{X_1, X_2, \dots, X_n\}$ dont la loi dépend d'un paramètre θ.

Soit $\hat{\theta}_n$ l'estimateur du maximum de vraisemblance de θ.

Alors $\hat{\theta}_n$ est asymptotiquement normal, et on a :

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow{n\to\infty} \mathcal{N}\left(0, \frac{1}{\sqrt{I_1(\theta)}}\right)$$

où $I_1(\theta)$ désigne l'information de Fisher

On en déduit que l'estimateur $\hat{\theta}_n$ est **asymptotiquement**

- **sans biais**
- **efficace**

$$\hat{\theta}_n \xrightarrow{n\to\infty} \mathcal{N}\left(\theta, \frac{1}{\sqrt{I_n(\theta)}}\right)$$

Méthode Delta : loi d'une transformée d'une V.A. asymptotiquement normale

Objectif

- Obtenir une approximation de la distribution asymptotique de la transformée d'une V.A. asymptotiquement normale
- Plus généralement, on peut considérer la méthode delta comme une **extension du TCL** (théorème central limite)

Énoncé

Soit $\hat{\theta}_n$ un estimateur asymptotiquement normal de θ :

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow{n\to\infty} \mathcal{N}\big(0, J(\theta)\big)$$

Pour toute fonction g dérivable telle que $g'(\theta) \neq 0$:

$$\sqrt{n}\left(g(\hat{\theta}_n) - g(\theta)\right) \xrightarrow{n\to\infty} \mathcal{N}(0, J(\theta)[g'(\theta)]^2)$$

Exemple

- Déterminer la distribution asymptotique de $Y = \exp(\bar{X})$

4. Estimation par intervalle de confiance

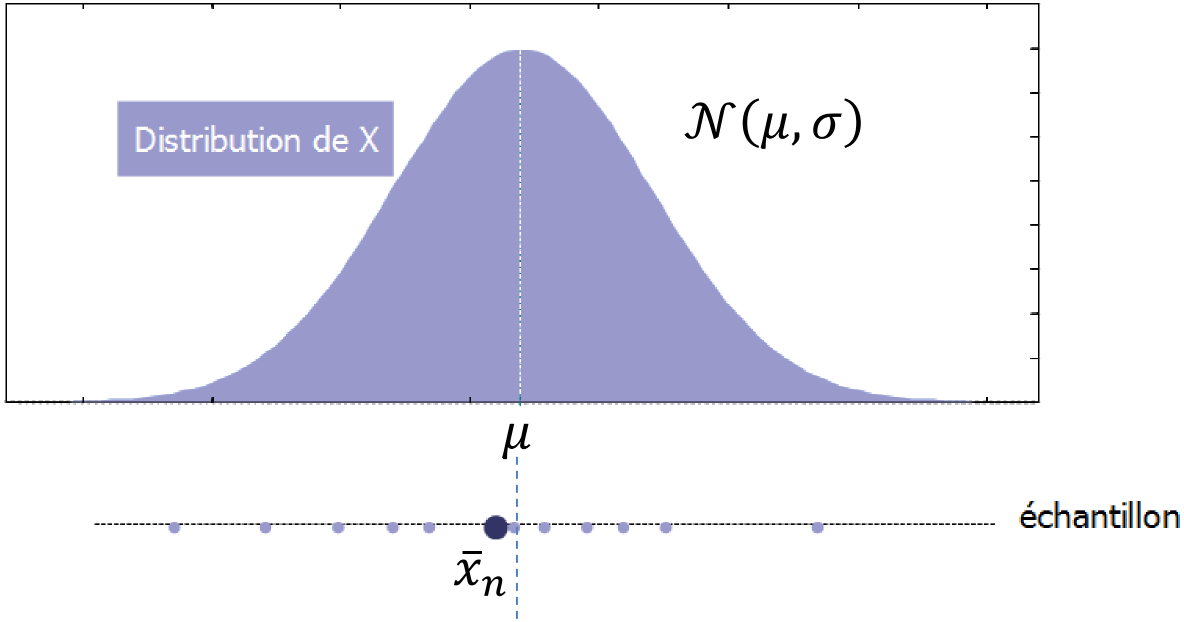
4.1 – Intervalle de confiance d'une moyenne

Introduction « intuitive »

- Soit un échantillon aléatoire simple $\{X_1, X_2, \ldots, X_n\}$ de taille n issu d'une population gaussienne de moyenne μ et de variance σ^2
- On sait que $\bar{X}_n$ est un « bon » estimateur de la moyenne μ de la population
- Sur un échantillon donné, l'estimation $\bar{x}_n$ ne coïncide pas avec la valeur de μ ...

4. Intervalle de confiance – IC d'une moyenne

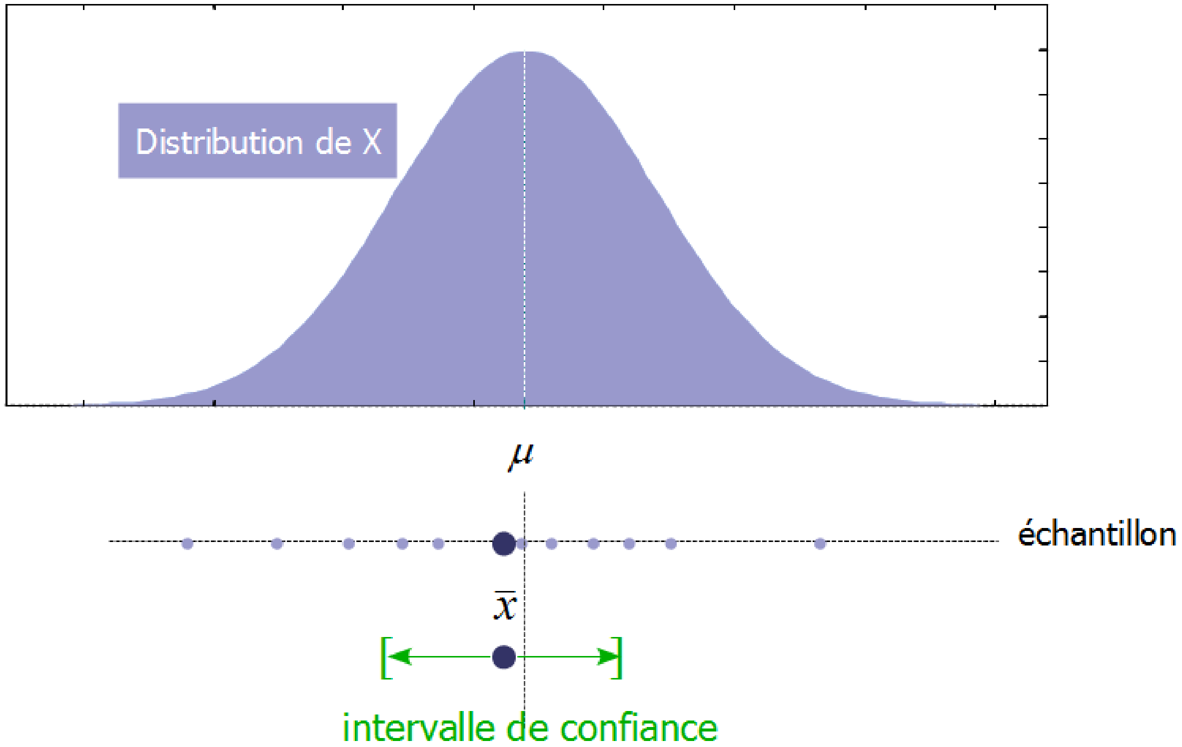
Illustration



Objectif de l'Intervalle de Confiance (IC)

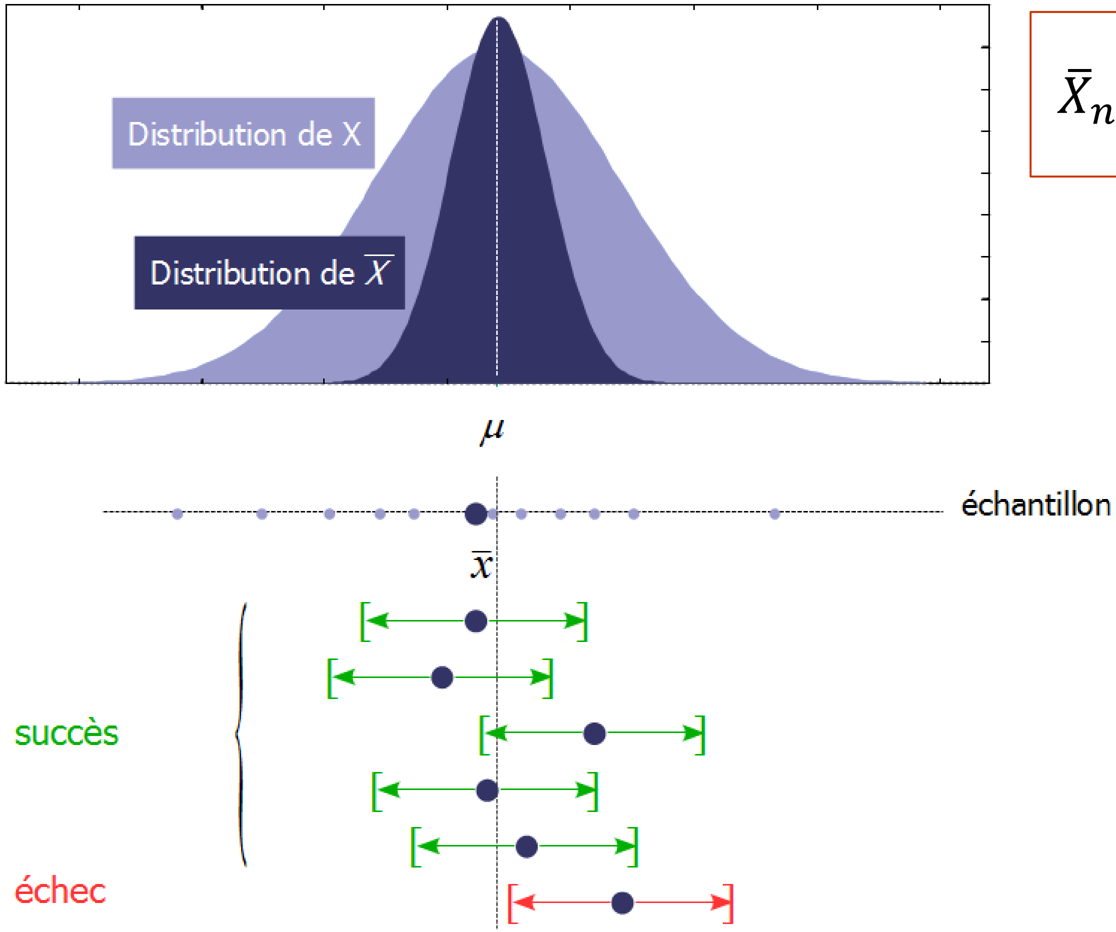
- Fournir un intervalle (une « **fourchette d'estimation** ») qui possède une forte probabilité d'encadre la vraie valeur du paramètre μ
- On dira aussi que l'on a une **forte confiance** dans le fait que notre intervalle encadre la vraie valeur du paramètre

4. Intervalle de confiance – IC d'une moyenne



Idée : ajouter une « marge d'erreur » de façon à encadrer μ

Quelles sont les valeurs attendues pour $\bar{X}_n$?



$$\bar{X}_n \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Selon l'échantillon tiré, **pour une marge donnée**, l'intervalle parviendra à encadrer μ ou échouera…

Construction de la marge d'erreur

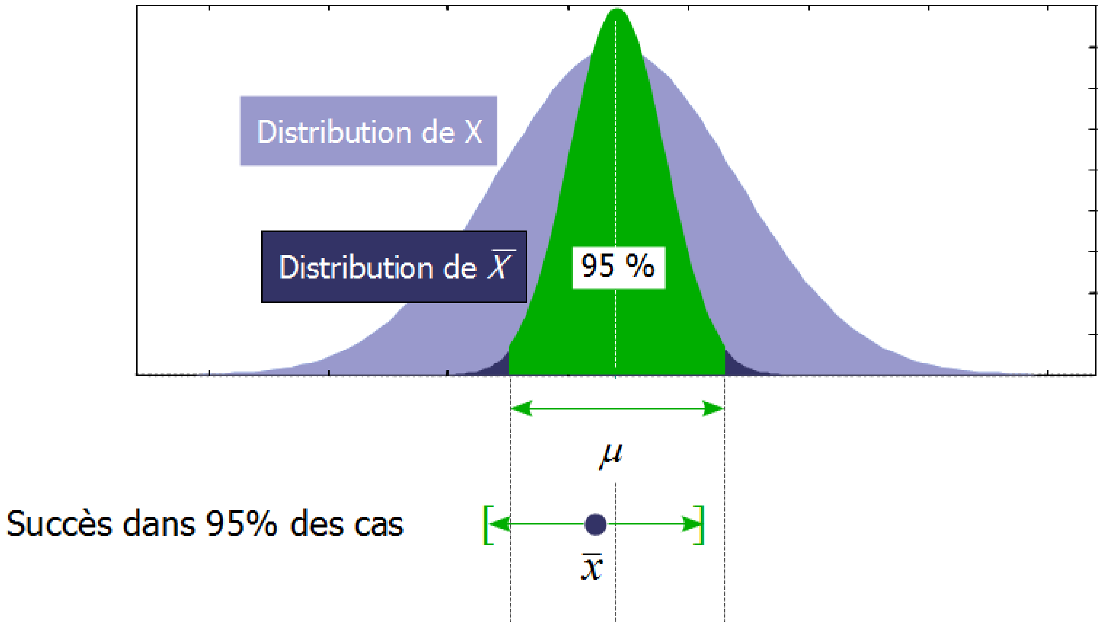
Deux objectifs antagonistes :

- Un intervalle de faible amplitude = assez précis
- Un intervalle d'amplitude assez importante pour encadrer souvent μ

En pratique

On **fixe la confiance** avec laquelle on souhaite travailler,
C'est-à-dire : on fixe **le taux de succès** avec lequel l'intervalle atteindra son but

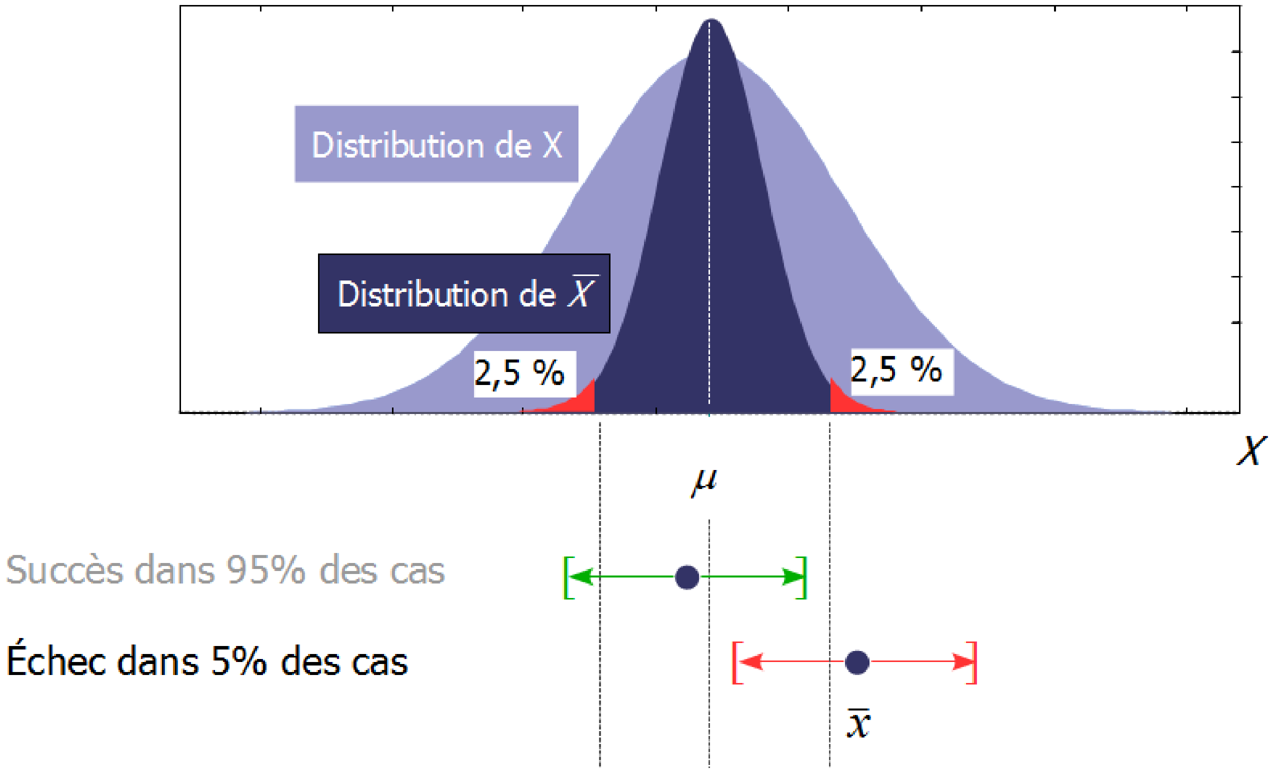
Illustration pour une confiance de 95%



La vrai valeur μ est correctement encadrée pour **95% des échantillons** issus de la population

Notre intervalle contient la vraie valeur de μ avec une **confiance de 95%**

4. Intervalle de confiance – IC d'une moyenne



Les **5% des échantillons**, associés aux valeurs de $\bar{x}_n$ les plus extrêmes, fournissent un intervalle de confiance qui n'encadre pas μ

Notre intervalle échoue à encadrer la vraie valeur de μ avec **un risque de 5%**

Définition d'un IC pour μ au niveau $(1-\alpha)$

Soit un échantillon aléatoire simple $\{X_1, X_2, \ldots, X_n\}$ de taille n issu d'une population gaussienne de moyenne μ et de variance σ^2

On appelle intervalle de confiance pour la moyenne μ au seuil $(1-\alpha)$ l'intervalle $\mathrm{IC}_{(1-\alpha)}$ qui a la probabilité $(1-\alpha)$ de contenir la vraie valeur de μ :

$$\boxed{P\left(\mu \in \mathrm{IC}_{(1-\alpha)}\right) = (1-\alpha)}$$

Construction de l'intervalle

$$\boxed{\mathrm{IC}_{(1-\alpha)}(\mu) = \left[\bar{X}_n - u_{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}} \,;\, \bar{X}_n + u_{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}\right]}$$

IC d'une moyenne μ dans le cas général

Il dépend de la **distribution de la moyenne d'échantillon $\overline{X}_n$**

	Loi $(X) = \mathcal{N}$		Loi (X) quelconque	
	σ connu	σ inconnu	σ connu	σ inconnu
$n \geq 30$	$\sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$	$\sim T\ (n-1)$ ou TCL	$\approx \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$	$\approx \mathcal{N}\left(\mu, \frac{s_c}{\sqrt{n}}\right)$
$n < 30$	$\sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$	$\sim T(n-1)$	?	?

On distingue usuellement les cas suivants :

- L'échantillon est de « grande » taille : application du TCL
- L'échantillon est de « petite » taille, X n'est pas normale
- L'échantillon est de « petite » taille, X est normale et σ^2 est connue
- L'échantillon est de « petite » taille, X est normale mais σ^2 est inconnue

4.2 – Intervalle de confiance – Cas général

Définition d'un IC pour θ au niveau $(1-\alpha)$

Soit un échantillon aléatoire simple $\{X_1, X_2, \ldots, X_n\}$ de taille n issu d'une population où X suit une loi de probabilité P_θ dépend d'un paramètre θ

On appelle intervalle de confiance pour la moyenne θ au seuil $(1-\alpha)$ l'intervalle $\mathrm{IC}_{(1-\alpha)}$ qui a la probabilité $(1-\alpha)$ de contenir la vraie valeur de θ

$$\boxed{P\left(\theta \in \mathrm{IC}_{(1-\alpha)}\right) = (1-\alpha)}$$

Statistique libre ou asymptotiquement libre

On appelle statistique (asymptotiquement) libre pour θ, toute statistique dont la loi (asymptotique) ne dépend pas du paramètre θ

Exemples. Soit un échantillon aléatoire simple $\{X_1, X_2, \dots, X_n\}$ de taille n issu d'une population gaussienne de moyenne μ et de variance σ^2

- Si σ^2 est connue, $\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$ est libre pour μ, de loi $N(0,1)$
- Si σ^2 est inconnue, $\frac{\bar{X}_n - \mu}{\frac{S_C}{\sqrt{n}}}$ est libre pour μ, de loi de Student $T(n-1)$
- Si μ est connue, $\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2$ est libre pour σ^2, de loi $\chi^2(n)$
- Si μ est inconnue, $\sum_{i=1}^{n} \left(\frac{X_i - \bar{X}_n}{\sigma}\right)^2$ est libre pour σ^2, de loi $\chi^2(n-1)$

Construction « pratique » d'un intervalle de confiance

Soit T_n un estimateur du paramètre θ

On cherche $\varphi_1(T_n)$ et $\varphi_2(T_n)$ expressions indépendantes de θ et telles que :

$$\boxed{P(\varphi_1(T_n) \leq \theta \leq \varphi_2(T_n)) = 1 - \alpha}$$

Démarche. Illustration pour la construction d'un IC pour μ, σ connu

- Estimateur de μ : $\bar{X}_n$
- Statistique libre pour μ : $\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1)$
- Encadrement pour la statistique libre : $P\left(-u_{1-\frac{\alpha}{2}} \leq \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \leq u_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$
- On en déduit l'encadrement pour μ : $P\left(\bar{X}_n - u_{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + u_{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$

Remarque

Le choix d'une statistique libre permet d'éviter que les bornes de l'intervalle de confiance dépendent du paramètre θ

Application aux quatre exemples précédents

1°) $$\mathrm{IC}_{(1-\alpha)}(\mu) = \left[\bar{X}_n - u_{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}} \;\; ; \;\; \bar{X}_n + u_{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}\right]$$

2°) $$\mathrm{IC}_{(1-\alpha)}(\mu) = \left[\bar{X}_n - t_{1-\frac{\alpha}{2}}(n-1) \times \frac{S_c}{\sqrt{n}} \;\; ; \;\; \bar{X}_n + t_{1-\frac{\alpha}{2}}(n-1) \times \frac{S_c}{\sqrt{n}}\right]$$

3°) $$\mathrm{IC}_{(1-\alpha)}(\sigma^2) = \left[\frac{n{S^*}^2}{k_{1-\frac{\alpha}{2}}(n)} \;\; ; \;\; \frac{n{S^*}^2}{k_{\frac{\alpha}{2}}(n)}\right]$$

4°) $$\mathrm{IC}_{(1-\alpha)}(\sigma^2) = \left[\frac{(n-1){S_c}^2}{k_{1-\frac{\alpha}{2}}(n-1)} \;\; ; \;\; \frac{(n-1){S_c}^2}{k_{\frac{\alpha}{2}}(n-1)}\right]$$

Intervalle de confiance asymptotique

- On construit l'intervalle de confiance pour θ en utilisant l'estimateur du maximum de vraisemblance $\hat{\theta}_n$ de θ
- On sait que cet estimateur est asymptotiquement normal, donc :

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow{n\to\infty} \mathcal{N}\left(0, \frac{1}{\sqrt{I_1(\theta)}}\right)$$

- On en déduit l'intervalle asymptotique suivant pour θ :

$$\boxed{\mathrm{IC}_{(1-\alpha)}(\theta) = \left[\hat{\theta}_n - \frac{u_{1-\frac{\alpha}{2}}}{\sqrt{nI_1(\theta)}} \; ; \hat{\theta}_n + \frac{u_{1-\frac{\alpha}{2}}}{\sqrt{nI_1(\theta)}}\right]}$$

En pratique, on remplace $I_1(\theta)$ par $I_1\left(\hat{\theta}_n\right)$

5. Tests d'hypothèses

« Un **test d'hypothèse** est une démarche consistant à évaluer une hypothèse (portant sur une population étudiée) en fonction de données observées dans un échantillon »

Plan

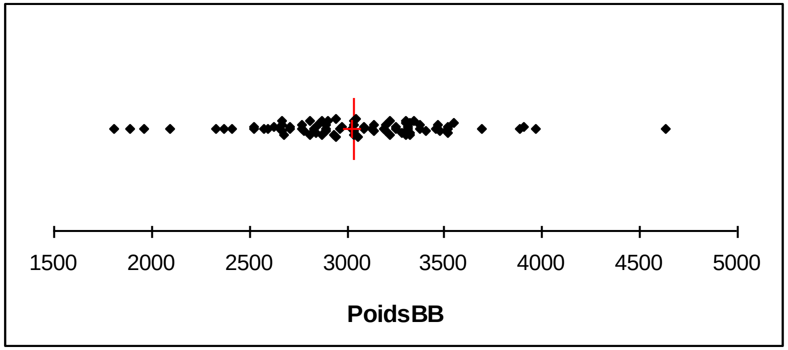
- Exemple introductif : conformité d'une moyenne
- Généralités sur les tests paramétriques
- Approche de Neyman et Pearson
- Tests d'adéquation
- Tests d'indépendance
- Test d'ajustement de Kolmogorov - Smirnov

5.1 – Exemple introductif

Exemple

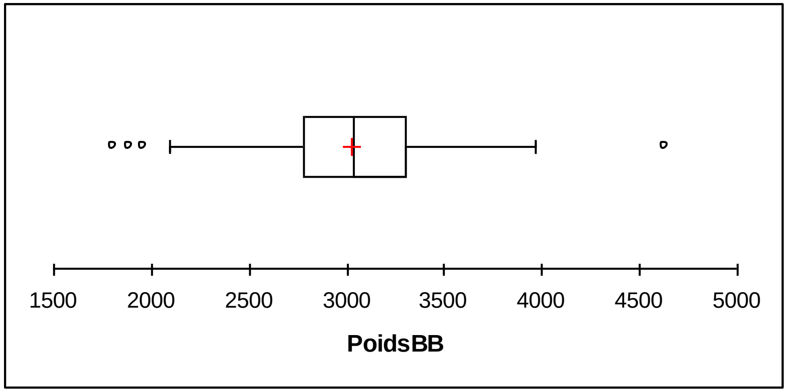
On étudie le poids moyen d'un BB dans une maternité. Est-il ou non identique à celui de la population française (poids moyen = 3250 g)

On dispose d'un échantillon du poids (X) de 83 bébés nés dans cette maternité



Questions

- Le poids moyen d'un bébé dans cette maternité est-il **égal à 3250 g** ? (est-elle *représentative* de la pop. fra. ?)
- Le poids moyen d'un bébé dans cette maternité est-il **inférieur à 3250** ?



Le déroulement du test

Il suit en général les étapes suivantes :

- **Hypothèses testées H_0 et H_1**
  Questions que l'on se pose, au niveau de la population étudiée
  Elles s'expriment toujours sous la forme de deux hypothèses
- **Choix d'une statistique de test** (ou de décision) et **sa loi sous H_0**
  C'est une quantité ou indicateur que l'on construit à partir
  des données de l'échantillon et qui nous aide à fonder notre décision
- **Définition de la région de rejet** de l'hypothèse centrale testée
  On confronte la valeur de la statistique de test à cette région
- **Règle de décision – Conclusion**
  On retient l'une ou l'autre des hypothèses testée,
  avec un certain risque d'erreur

Hypothèses testées

- H_0 : « Le poids moyen d'un bébé μ = 3250 g » **Hypothèse « nulle »**
- H_1 : « Le poids moyen d'un bébé μ < 3250 g » **Hypothèse « alternative »**

Rôle central de l'hypothèse H_0

- L'hypothèse nulle est **l'hypothèse de référence**
  C'est elle qui est soumise à l'épreuve des données

« Les données observées dans l'échantillon sont-elles **en accord avec l'hypothèse H_0** ?

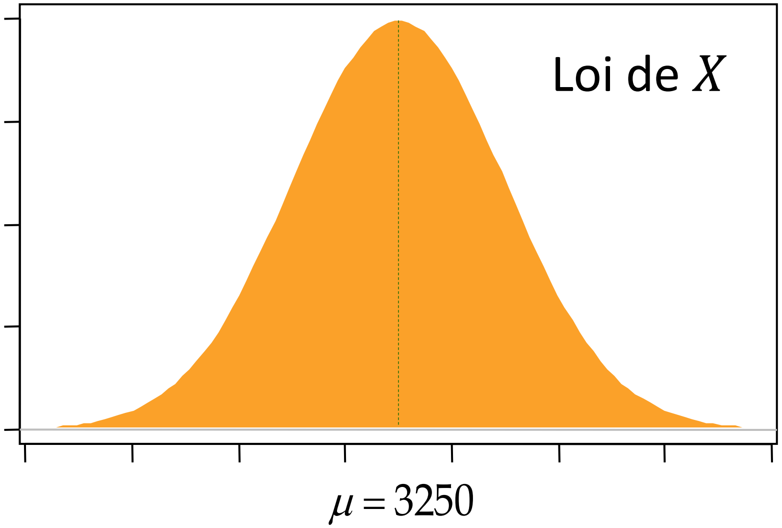
- Elle exprime souvent l'absence de différence, de relation, le « fruit du hasard », etc.

5. Tests d'hypothèses – Exemple introductif

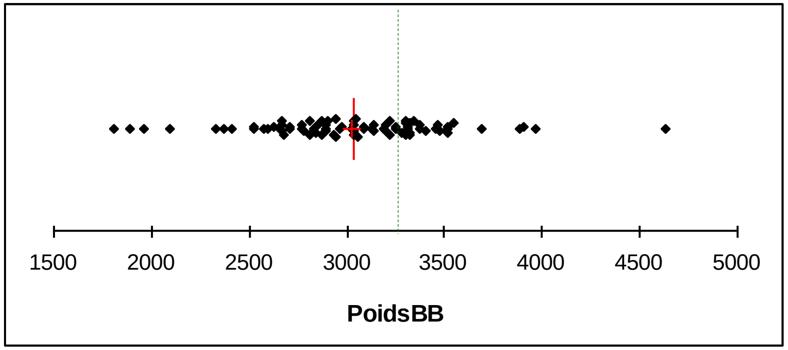
Question

- Est-il vraisemblable / plausible
  que les 83 observations proviennent
  d'une population de moyenne $\mu = 3250$?
  Hyp. : loi de X supposée inconnue,
  (représentée ici par une normale dans
  un souci de simplification)

- Quelles sont **les chances** qu'un tel
  échantillon soit le résultat d'un
  tirage au hasard dans une population
  de moyenne $\mu = 3250$?



Tirage au hasard **de 83 observations ?**



L'hypothèse alternative H_1 – Test bilatéral ou unilatéral

- C'est l'hypothèse (une famille d'hypothèses en général !) vers laquelle on se tournera en cas de rejet de l'hypothèse nulle

- L'hypothèse alternative peut être de deux types :

  - H_1 : « Le poids moyen d'un bébé < 3250 g »
  - H_1 : « Le poids moyen d'un bébé > 3250 g »

  Test unilatéral

  - H_1 : « Le poids moyen d'un bébé ≠ 3250 g »

  Test bilatéral

- Choix d'un test bilatéral : on a ***pas d'idée a priori*** sur le sens de la différence

Statistique de test

- C'est un indicateur calculé à partir des données de l'échantillon
- C'est sur la base de cette statistique que l'on évalue le caractère vraisemblable de l'hypothèse H_0
- Le choix de la statistique de test directement naturellement de l'expression de l'hypothèse H_0

Test de la moyenne μ d'une population → **Statistique de test = moyenne $\overline{X}_n$ de l'échantillon**

Test de la proportion π d'une population → Statistique de test = proportion p de l'échantillon

Test de la variance σ^2 d'une population → Statistique de test = variance s^2 de l'échantillon

5. Tests d'hypothèses – Exemple introductif

Statistique de test

- C'est un indicateur calculé à partir des données de l'échantillon
- C'est sur la base de cette statistique que l'on évalue le caractère vraisemblable de l'hypothèse H_0
- Le choix de la statistique de test directement naturellement de l'expression de l'hypothèse H_0

Test de la moyenne μ d'une population → **Statistique de test = moyenne $\overline{X}_n$ de l'échantillon**

Valeur de la statistique dans l'échantillon $\overline{x} = 3025$

« En tirant un échantillon de taille n = 83 dans une population de moyenne 3250, est-il vraisemblable d'obtenir une telle moyenne (aussi faible) ?

Les valeurs attendues de $\bar{X}_n$ sous H_0

Or, lorsque l'on prélève un échantillon de taille *n* dans une population de moyenne μ = 3250, on sait à quelles valeurs de $\bar{X}_n$ on doit s'attendre (*cf* loi d'une moyenne d'échantillon, p. 140)

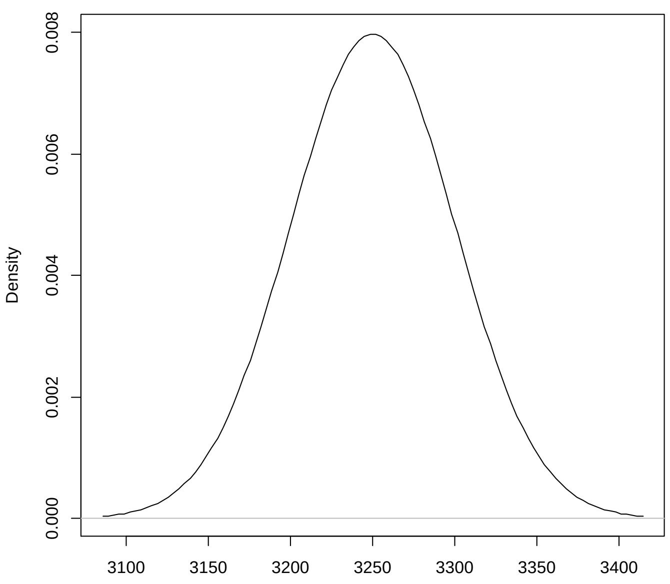
Loi de $\bar{X}_n$: $n \geq 30$, on applique le TCL

$$\bar{X}_n \approx \mathcal{N}\left(\mu = 3250, \frac{s_c}{\sqrt{n}} = \frac{455{,}74}{\sqrt{83}} = 50{,}02\right)$$

où s_c est l'estimation non biaisée de l'écart type de la population, calculée à partir des données de l'échantillon

Cette loi de $\bar{X}_n$ est appelée
loi de la statistique de test sous *H_0*

$\bar{X}_n \approx \mathcal{N}(3250\,; 50{,}02)$

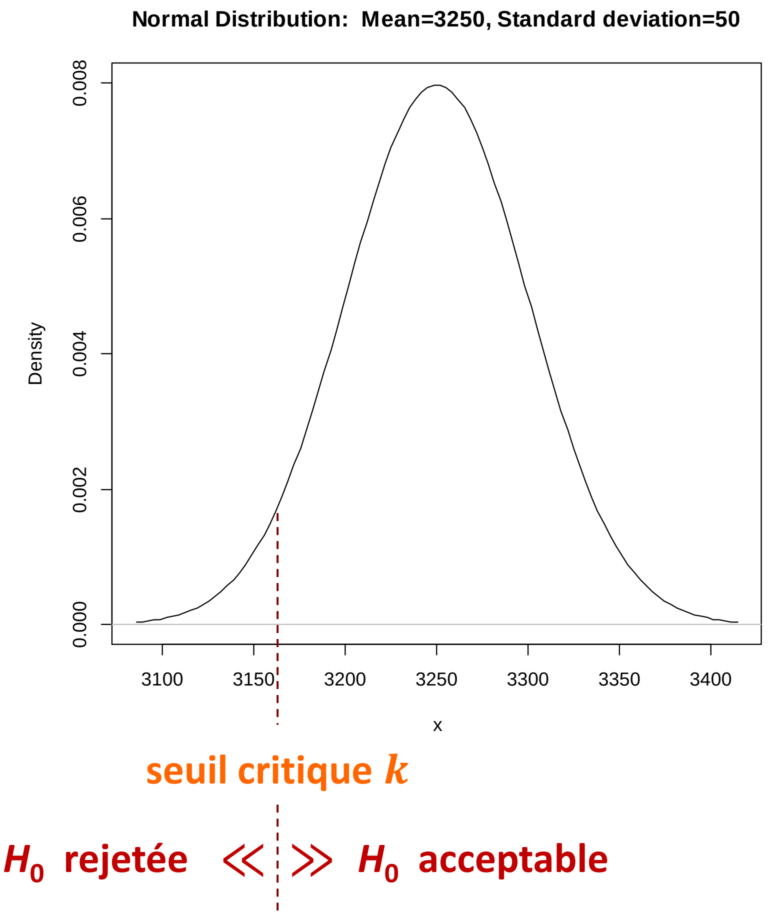


Région de rejet de H_0

Tant que la moyenne reste « proche » de μ = 3250, on considèrera que l'hypothèse testée H_0 reste plausible ou vraisemblable

Que veut dire « proche » ?

Nécessité de définir un **seuil critique k** en deçà duquel on décidera au contraire que la moyenne d'échantillon observée n'est plus compatible avec l'hypothèse H_0



5. Tests d'hypothèses – Exemple introductif

Région de rejet de H_0

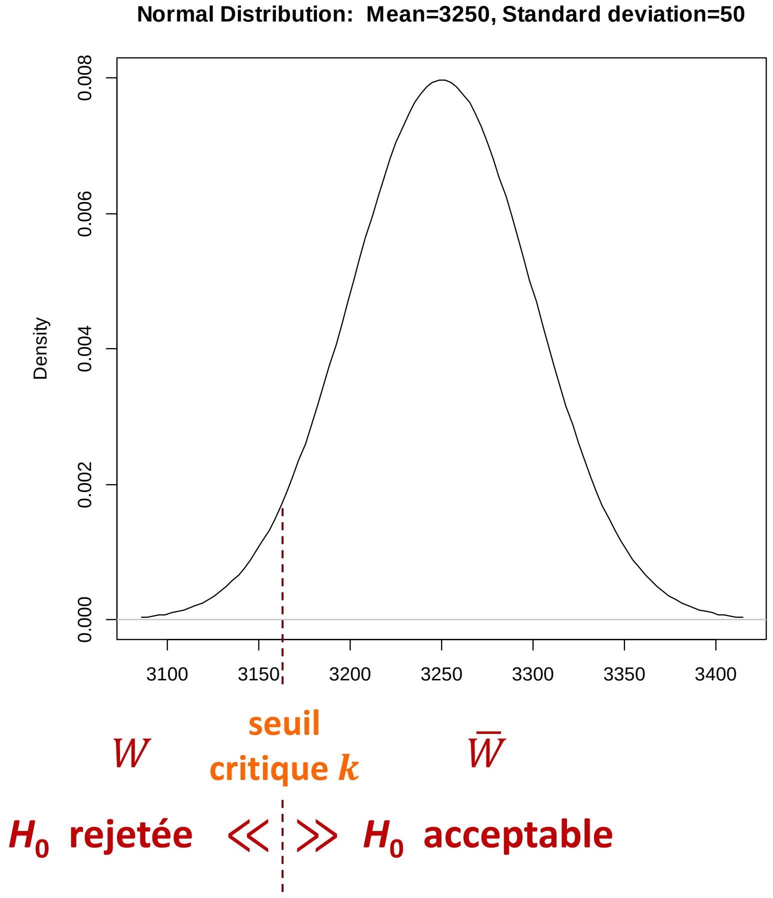
On définit ainsi la région de rejet de H_0

$$\boxed{W =]-\infty \,;k]}$$

Avec la règle suivante :

Si $\bar{X}_n \in W$: **H_0 rejetée**
Si $\bar{X}_n \notin W$: **H_0 acceptable**

W = ensemble des valeurs de la statistique de test qui conduit à rejeter H_0



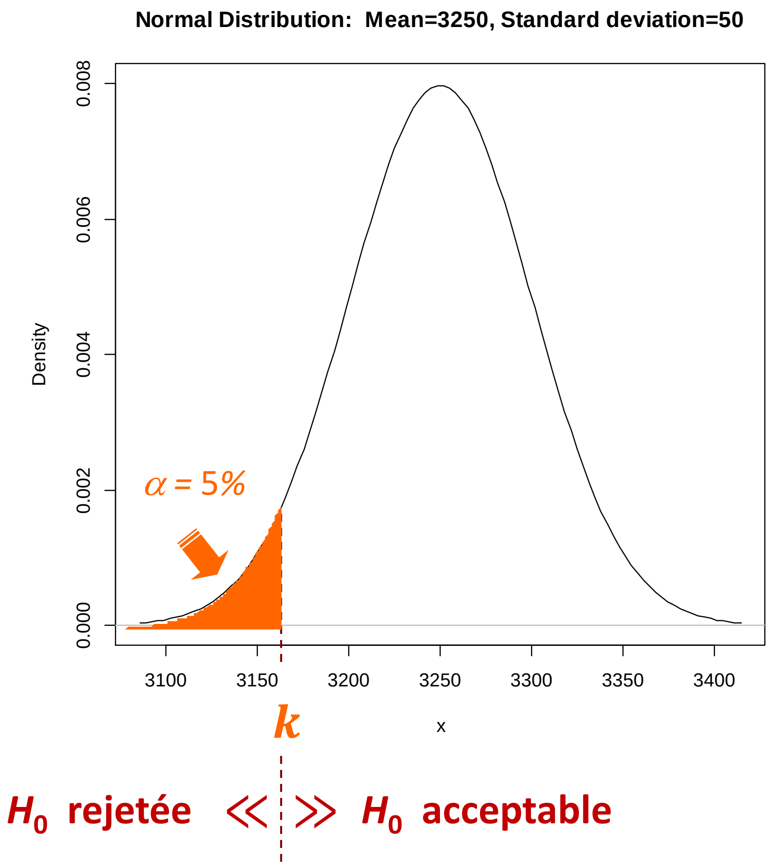
Comment déterminer la valeur du seuil k ?

Seuil critique k : le choix du **risque de première espèce α**

- Supposons que l'on décide de rejeter H_0 si la moyenne $\bar{X}_n$ observée fait partie des $\alpha = 5\%$ plus petites valeurs de la moyenne
- Ce faisant, on prend un risque ! Car la probabilité de voir apparaitre une moyenne aussi peu élevée sous H_0 n'est pas nulle : elle vaut 5%
- Notre risque d'erreur de se tromper en rejetant H_0 est donc précisément égal à $\alpha = 5\%$

$$\boxed{\alpha = P(\text{rejet } H_0 \mid H_0 \text{vraie})}$$

$\alpha =$ risque de rejet à tort de H_0
$\alpha =$ risque de première espèce



Calcul pratique du seuil critique *k*

Il est calculé en exprimant le risque α

$$P(\bar{X}_n \leq k | H_0) = \alpha = 5\%$$

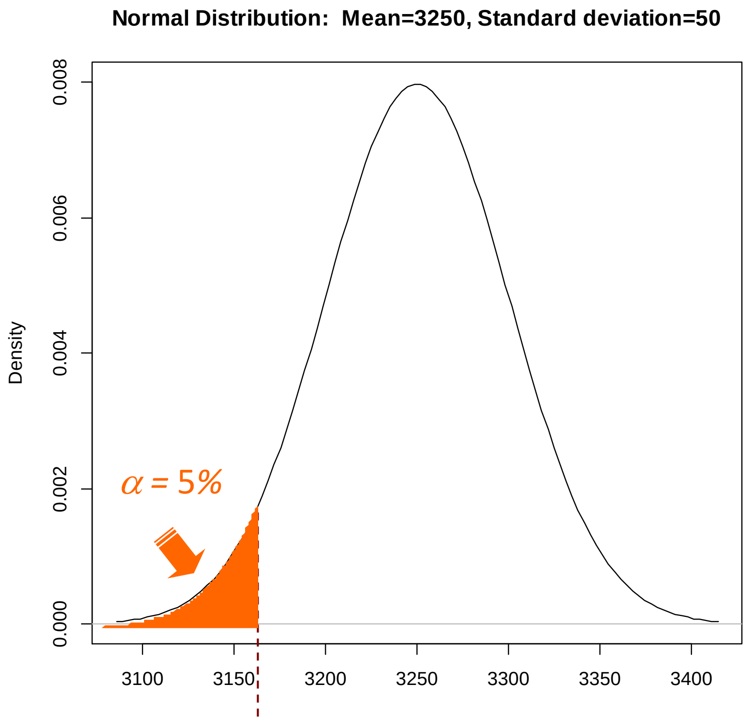
Or, sous l'hypothèse H_0

$$\bar{X}_n \approx \mathcal{N}(3250\,;\ 50{,}02)$$

On en déduit

$$P\left(U \leq \frac{k - 3250}{50{,}02}\right) = 0{,}05$$

$$\boldsymbol{k = 3168}$$



$$\boldsymbol{k = 3168}$$

***H_0* rejetée ≪ ≫ *H_0* acceptable**

Règle de décision – Conclusion

- **Si** $\bar{X}_n \in W$

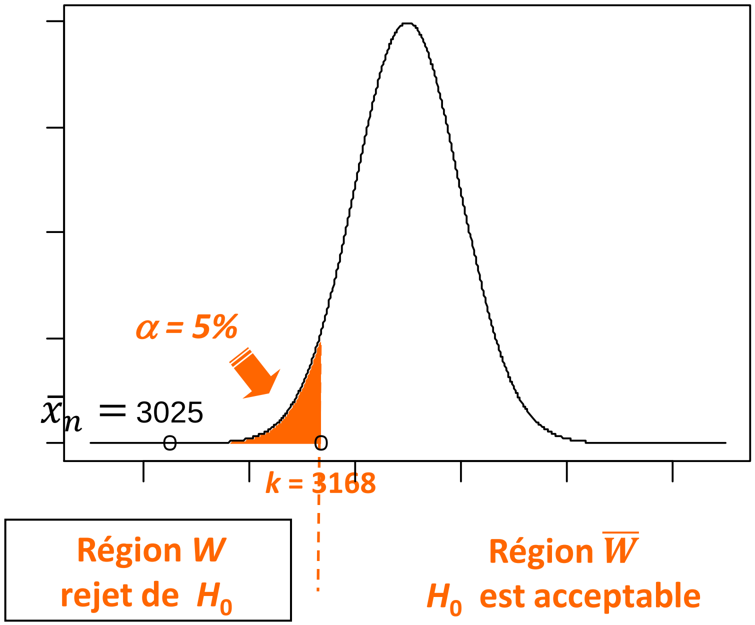
  La moyenne observée a peu de chance d'apparaître sous l'hypothèse H_0

  Rejet de H_0

- **Si** $\bar{X}_n \notin W$

  La moyenne observée a une probabilité « non négligeable » d'apparaître sous H_0

  H_0 est acceptable



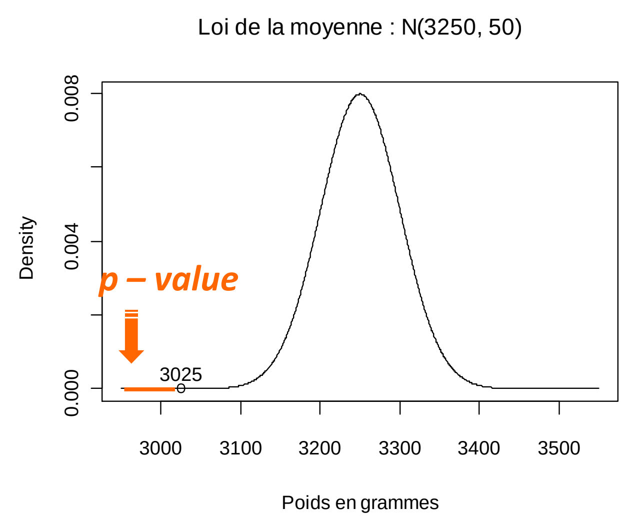
Application : $\bar{x}_n = 3025 \in W$: H_0 : $\mu = 3250$ g **est donc rejetée,**
au seuil de risque $\alpha = 5\%$

Règle de décision basée sur la *p – value*

- Règle de décision la plus fréquemment utilisée depuis le développement des logiciels statistiques
- On calcule la probabilité d'obtenir une valeur de la statistique de test **aussi « extrême »** que celle observée dans l'échantillon, lorsque H_0 est vraie

$$p - value = P(\bar{X} \leq 3025 | H_0 \text{ vraie})$$

$$= \boxed{3.4\ 10^{-6}}$$



5. Tests d'hypothèses – Exemple introductif

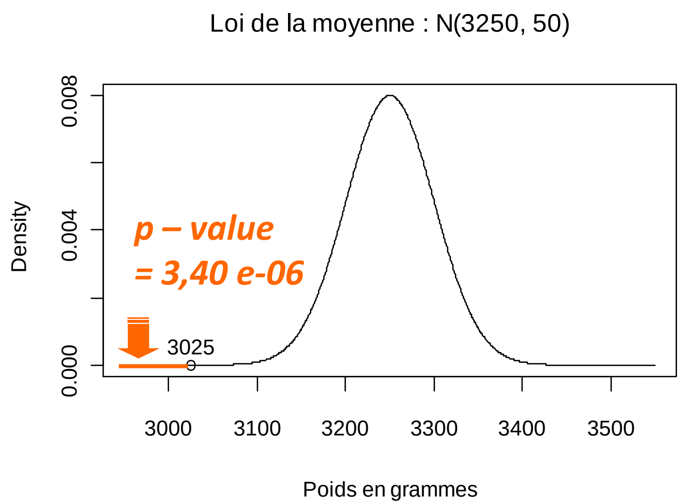
- Il est **extrêmement rare** d'obtenir une moyenne aussi faible par un simple tirage au hasard dans une population où le poids moyen = 3250
- Les données de notre échantillon sont **très peu compatibles avec H_0**
- Nous pouvons donc « raisonnablement » **rejeter l'hypothèse H_0**

Cette décision est-elle risquée ?

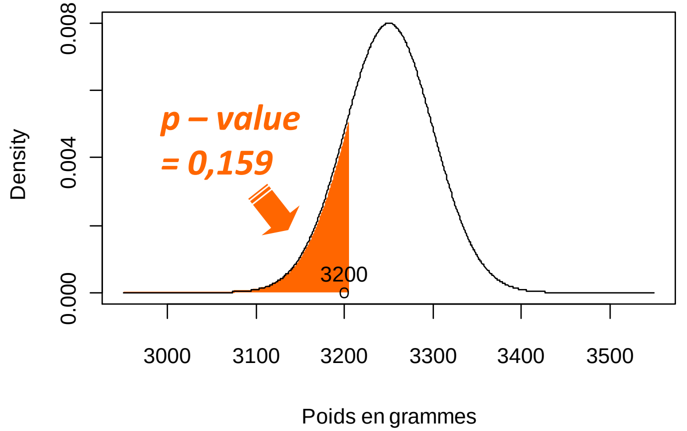
- Lorsqu'on rejette H_0, le **risque réellement encouru de se tromper** est mesuré par la *p – value !*

5. Tests d'hypothèses – Exemple introductif

- En rejetant H_0 pour une moyenne = 3025, on ne prend quasiment aucun risque (p = 3,40 e-06)



- En rejetant H_0 pour une moyenne = 3200, on prend un risque assez important (p = 0,159)



Le niveau de risque que l'on est prêt à accepter au maximum =
le risque de première espèce α

La règle de décision usuelle dans un test d'hypothèse

Comparaison de la $p-value$ au risque α

- **Si** $\boldsymbol{p-value < \alpha}$: Le risque encouru en rejetant H_0 est faible par rapport au risque « acceptable »

  On décide de rejeter H_0

- **Si** $\boldsymbol{p-value > \alpha}$: Le risque encouru en rejetant H_0 est trop élevé par rapport au risque « acceptable »

  L'hypothèse H_0 est acceptable

Equivalence avec « l'ancienne » règle de décision

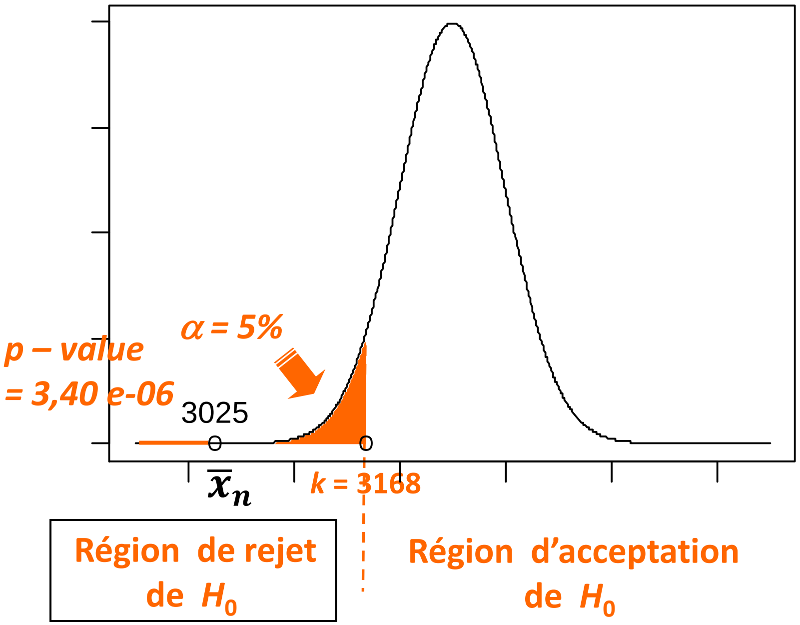
Comparaison de la statistique de test à la région de rejet

- $\overline{X}_n \in W \Leftrightarrow p - value < \alpha$

  Rejet de H_0

- $\overline{X}_n \notin W \Leftrightarrow p - value > \alpha$

  H_0 acceptable



5.2 – Généralités sur les tests paramétriques

L'enjeu d'un test d'hypothèse

- Un test d'hypothèse consiste à définir **une règle de décision** concernant la **validité d'une hypothèse** portant sur un (ou plusieurs) paramètres d'une (ou plusieurs) populations dont on observe un (ou plusieurs) échantillons aléatoires
- Deux hypothèses en présence :
  - H_0 est appelée hypothèse nulle (**hypothèse de référence** : rejet ou non rejet)
  - H_1 (ou H_A) est l'hypothèse alternative
- Très souvent : existence de relation, de différence, d'écart à une cible visée, etc. L'enjeu du test est de parvenir à les mettre en évidence (= à rejeter H_0)

Les différents types de test d'hypothèses

Soit un échantillon $\{X_1, X_2, \ldots, X_n\}$ dont la loi dépend d'un paramètre θ.
De façon générale, un test d'hypothèses paramétrique s'écrit :

$$\begin{cases} H_0 : & \theta_0 \in \Theta_0 \\ H_1 : & \theta_1 \in \Theta_1 \end{cases}$$

- **Test simple** $\begin{cases} H_0 : & \theta = \theta_0 \\ H_1 : & \theta = \theta_1 \end{cases}$
- **Test unilatéral** $\begin{cases} H_0 : & \theta = \theta_0 \\ H_1 : & \theta > \theta_0 \end{cases}$ ou $\begin{cases} H_0 : & \theta = \theta_0 \\ H_1 : & \theta < \theta_0 \end{cases}$
- **Test bilatéral** $\begin{cases} H_0 : & \theta = \theta_0 \\ H_1 : & \theta \neq \theta_0 \end{cases}$
- **Test composite** $\begin{cases} H_0 : & \theta \in [\theta_a ; \theta_b] \\ H_1 : & \theta \notin [\theta_a ; \theta_b] \end{cases}$ ou $\begin{cases} H_0 : & \theta \notin [\theta_a ; \theta_b] \\ H_1 : & \theta \in [\theta_a ; \theta_b] \end{cases}$

Exemples de tests bilatéraux

Conformité d'une moyenne

L'utilisateur n'a priori aucune certitude quant au sens de la différence vis-à-vis de la valeur testée sous H_0

H_0 : « Le poids moyen d'un bébé $\mu = 3250$ g »

H_1 : « Le poids moyen d'un bébé $\mu \neq 3250$ g »

Égalité de deux moyennes

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

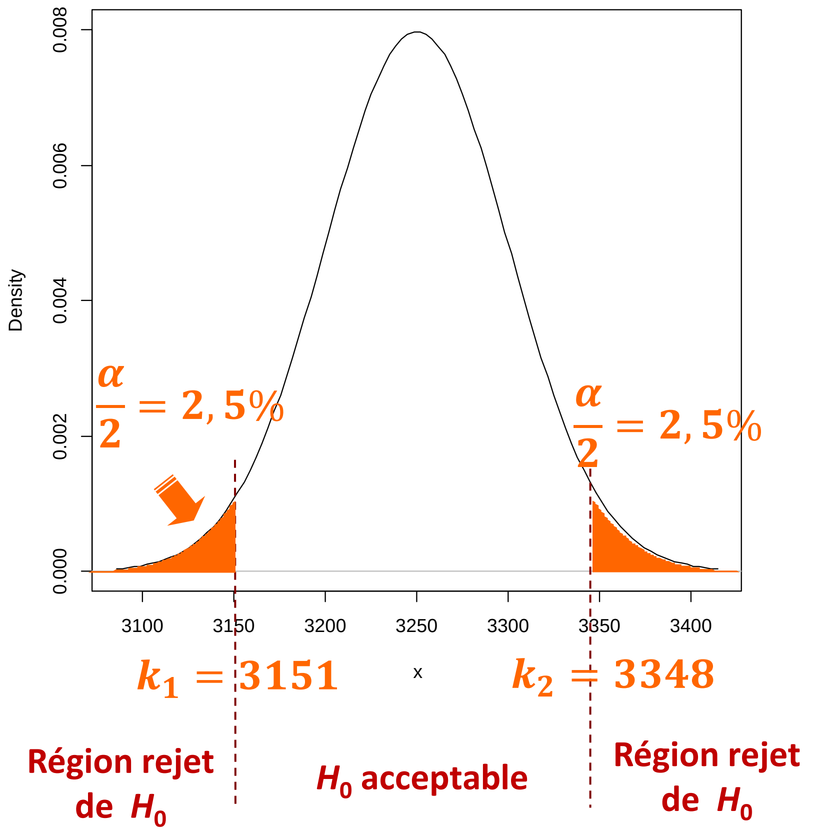
Conséquence (d'un test bilatéral) pour le déroulement du test

La région de rejet est maintenant formée de deux « sous-régions »

$$\boxed{W =]-\infty\,; k_1] \cup [k_2\,;\, +\infty[}$$

Les deux seuils de risque sont déterminés en fonction du risque α

$$P(\bar{X}_n \leq k_1 | H_0) = 2{,}5\%$$
$$P(\bar{X}_n \geq k_2 | H_0) = 2{,}5\%$$



Décisions possibles et probabilités associées

		Hypothèse « vraie »	
		H_0	H_1
Décision prise	H_0	Seuil de confiance $(1 - \alpha)$	Erreur de 2e espèce β
	H_1	Erreur de 1e espèce Seuil de signification α	Puissance du test $(1 - \beta)$

- **Risque de 1e espèce** α **:** *risque de rejet à tort de* H_0
- **Risque de 2e espèce** β **:** *risque d'acceptation à tort de* H_0
- **Seuil de confiance** $(1 - \alpha)$ **:** *acceptation à juste titre de* H_0
- **Puissance du test** $(1 - \beta)$ **:** *rejet à juste titre de* H_0

Exemple de calcul de risque de 2e espèce et de puissance de test

- Dans une autre maternité, on observe $\bar{x}_n = 3190$
- Quelle décision prend-on ?
- Calculer le risque de 2e espèce et la puissance du test dans l'hypothèse où le vrai poids moyen est égal à 3200 (puis pour 3100)
- Illustrer de manière graphique les quantités α , β et $(1 - \beta)$

Remarque : rôle joué par les hypothèses H_0 et H_1

- Le seuil critique *k* est déterminé sous l'hypothèse H_0 et pour un risque α

- Le risque β et la puissance $(1 - \beta)$ sont déterminés sous l'hypothèse H_1

5.3 – Approche de Neyman & Pearson

Principe

La méthode dite de *N&P* est une approche des tests d'hypothèses qui permet, pour **un risque α fixé**, de déterminer **la région optimale *W*** de rejet de l'hypothèse H_0, i.e. celle qui **maximise la puissance de test** $(1-\beta)$

Théorème

On considère le test d'hypothèses $\begin{cases} H_0 : & \theta = \theta_0 \\ H_1 : & \theta = \theta_1 \end{cases}$

Pour **un risque α fixé,** il existe un test optimal (de puissance $(1-\beta)$ maximale) dont la région de rejet *W* est définie par $\frac{L(x;\theta_1)}{L(x;\theta_0)} > c_\alpha$

où c_α est une constante dépendant du risque α

Interprétation

La région de rejet optimale de l'hypothèse H_0 (maximisant la puissance du test) comprend tous les échantillons tels leur vraisemblance sous l'hypothèse H_1 soit « plus importante » que celle sous H_0, pour un risque α fixé

Application

- Échantillon issu d'une loi gaussienne de variance connue
- Test simple de conformité d'une moyenne

$$\begin{cases} H_0 : & \mu = \mu_0 \\ H_1 : & \mu = \mu_1 \end{cases}$$

Avec $\mu_1 > \mu_0$

5.4 – Tests d'indépendance du χ^2

Exemple

- Échantillon de 3883 enfants d'un district écossais
- On dispose de deux informations : *sexe* et *couleur des cheveux*
- On forme le tableau croisé entre les **deux variables qualitatives** ou **tableau de contingence**

Variable qualitative 2

Variable qualitative 1

		couleur des cheveux					
		blond	*roux*	*châtain*	*brun*	*noir de jais*	
sexe	*garçon*	592	119	849	504	36	***2100***
	fille	544	97	677	451	14	***1783***
		1136	***216***	***1526***	***955***	***50***	***3883***

Problématique – Questions

- Les deux caractéristiques (sexe et couleur cheveux) sont-elles liées l'une à l'autre ?
- Existe-t-il un **lien de dépendance** entre les variables qualitatives ?
- La couleur des cheveux dépend-elle du sexe de l'enfant ?

Représentation graphique des effectifs du tableau de contingence

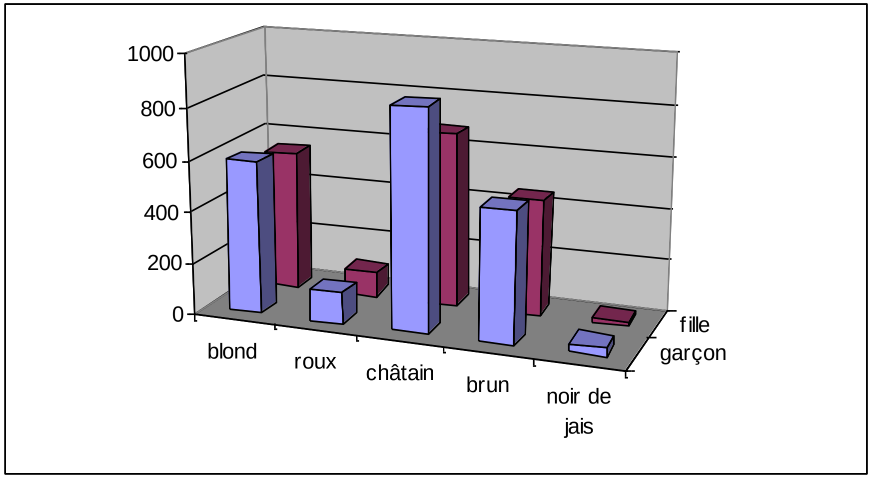


Tableau de contingence – Notations

- $n =$ nombre total d'individus de l'échantillon
- $n_{ij} =$ nombre d'individus possédant simultanément la modalité (i) de la variable 1 et la modalité (j) de la variable 2
- $n_{i\bullet} =$ nombre d'individus possédant la modalité (i) de la variable 1
- $n_{\bullet j} =$ nombre d'individus possédant la modalité (j) de la variable 2

	(1)	$\cdots$	(j)	$\cdots$	(J)	Total
(1)						$n_{1\bullet}$
			$\vdots$			$\vdots$
(i)		$\cdots$	n_{ij}	$\cdots$		$n_{i\bullet}$
			$\vdots$			$\vdots$
(I)						$n_{I\bullet}$
Total	$n_{\bullet 1}$	$\cdots$	$n_{\bullet j}$	$\cdots$	$n_{\bullet J}$	n

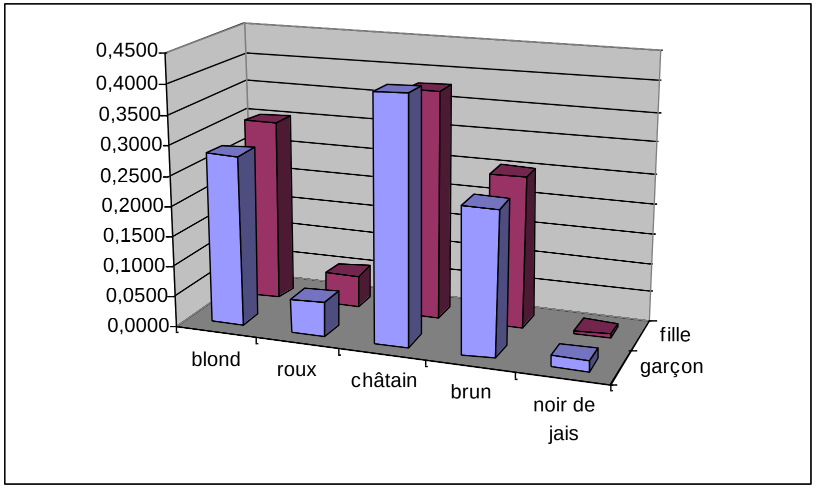
Profils lignes, profils colonnes et indépendance

Tableau des **profils lignes**

$$\left(\frac{n_{i1}}{n_{i\bullet}}, \dots, \frac{n_{ij}}{n_{i\bullet}}, \dots, \frac{n_{iJ}}{n_{i\bullet}}\right)$$

	blond	*roux*	*châtain*	*brun*	*noir de jais*	TOTAL
garçon	0,2819	0,0567	0,4043	0,2400	0,0171	1
fille	0,3051	0,0544	0,3797	0,2529	0,0079	1
	0,29	0,06	0,39	0,25	0,01	1

Graphique des **profils lignes**



Si la couleur des cheveux ne dépend pas du sexe, alors filles et garçons ont des profils de couleur de cheveux identiques

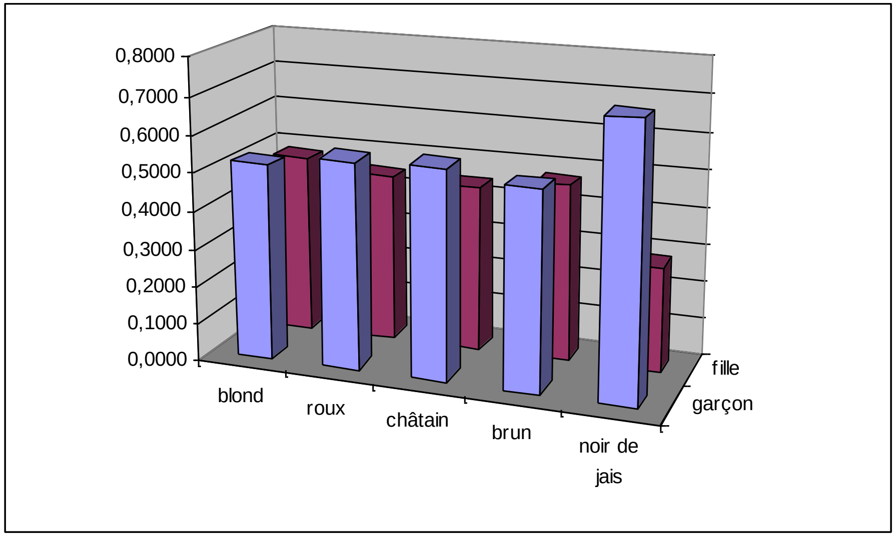
Profils lignes, profils colonnes et indépendance

Tableau des **profils colonnes**

$$\begin{pmatrix} \frac{n_{1j}}{n_{\bullet j}} \\ \vdots \\ \frac{n_{ij}}{n_{\bullet j}} \\ \vdots \\ \frac{n_{Ij}}{n_{\bullet j}} \end{pmatrix}$$

	blond	*roux*	*châtain*	*brun*	*noir de jais*	
garçon	0,5211	0,5509	0,5564	0,5277	0,7200	0,54
fille	0,4789	0,4491	0,4436	0,4723	0,2800	0,46
TOTAL	1	1	1	1	1	1

Graphique des **profils colonnes**



Si le sexe ne dépend pas de la couleur des cheveux, alors filles et garçons ont des profils de sexe identiques

Mise en œuvre du test du Chi2

Hypothèses testées

H_0 : *Couleur de cheveux* et *sexe* sont deux critères **indépendants**

H_1 : *Couleur de cheveux* et *sexe* **ne sont pas indépendants**

Hypothèses relatives à la population d'où est issu l'échantillon

Principe du test

Évaluer si les effectifs observés dans l'échantillon sont proches de ceux que l'on observerait en cas d'indépendance (= quand H_0 est vraie)

→ Quels sont les effectifs que l'on observerait en cas d'indépendance ?

Effectifs théoriques (t_{ij}) en cas d'indépendance

		couleur des cheveux					
		blond	roux	châtain	brun	noir de jais	
sexe	garçon	?	?	?	?	?	2100
	fille	?	?	?	?	?	1783
		1136	216	1526	955	50	3883

À marges constantes, comment se répartiraient les effectifs théoriques ?

En cas d'indépendance

	1	j	J	Σ
1				
i		t_{ij}		$n_{i\bullet}$
I				
Σ		$n_{\bullet j}$		n

Égalité des profils-lignes (ou colonne)

$$\frac{t_{ij}}{n_{i\bullet}} = \frac{n_{\bullet j}}{n} \Leftrightarrow t_{ij} = \frac{n_{i\bullet}\, n_{\bullet j}}{n}$$

Les effectifs théoriques en cas d'indépendance sont calculés par

$$t_{ij} = \frac{n_{i\bullet}\, n_{\bullet j}}{n}$$

5. Tests d'hypothèses – Test du Chi2 d'indépendance

Effectifs **observés**

n_{ij}

	blond	*roux*	*châtain*	*brun*	*noir de jais*	
garçon	592	119	849	504	36	2100
fille	544	97	677	451	14	1783
	1136	216	1526	955	50	3883

Effectifs **théoriques**

t_{ij}

	blond	*roux*	*châtain*	*brun*	*noir de jais*	
garçon	614,37	116,82	825,29	516,48	27,04	2100
fille	521,63	99,18	700,71	438,52	22,96	1783
	1136	216	1526	955	50	3883

Exemple $521{,}63 = \dfrac{1136 \times 1783}{3883}$

Les effectifs observés sont-ils très différents de ceux que l'on observerait en situation d'indépendance ?

➡ Nécessité de construire un **indicateur d'écart** entre ces deux tableaux

La statistique de décision (ou de test)

C'est une mesure globale de distance entre les deux tableaux

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

Application numérique : $D^2 = 10{,}47$

À quelles mesures de D^2 est-on en droit de s'attendre en cas d'indépendance (sous H_0) ?

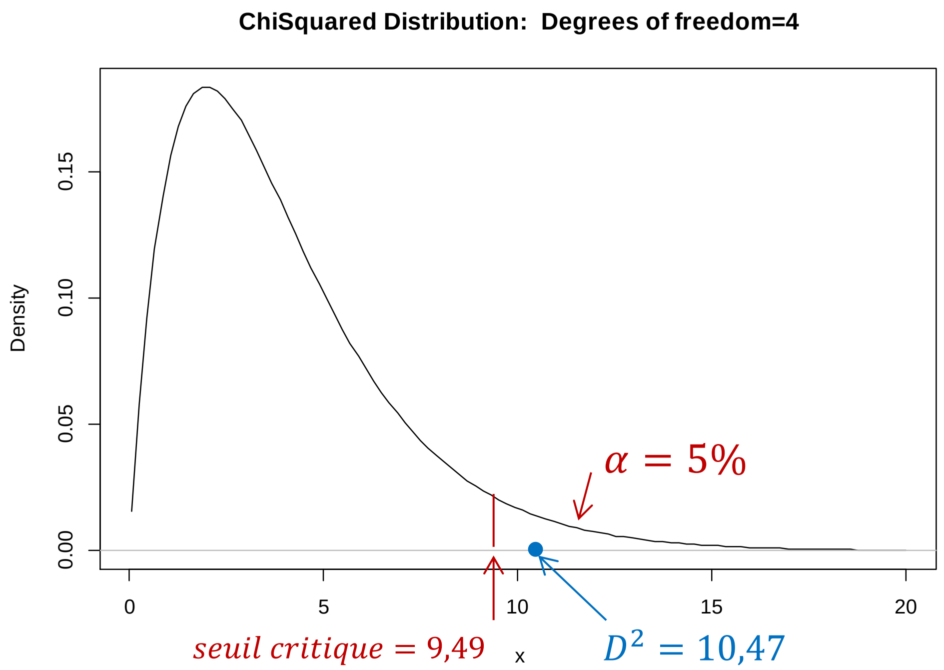
La loi de la statistique sous H_0

Quand H_0 est vraie, D^2 est distribuée selon une loi du Chi2 à $(I-1) \times (J-1)$ degrés de liberté

$$\boxed{D^2 \sim \chi_{(I-1)\,(J-1)}}$$

La règle de décision basée sur la région de rejet *W*

On rejette H_0 si la valeur de la statistique observée D^2 fait partie des α $(= 5\%$ par exemple) des valeurs les plus extrêmes auxquelles on peut s'attendre sous H_0



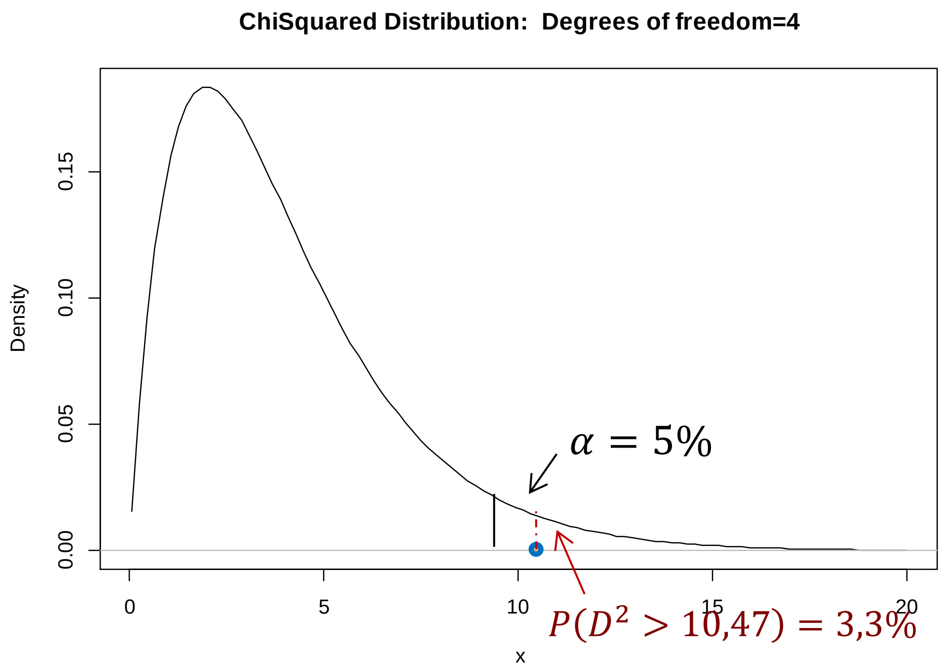
Conclusion

$D^2 = 10{,}47 \in W$

Au seuil de risque de 5%,
on peut rejeter l'hypothèse H_0

La règle de décision basée sur la probabilité critique (p – value)

On rejette H_0 si le risque réellement encouru de se tromper en rejetant H_0 au vu de la valeur de D^2 est inférieur au risque maximum que l'on est prêt à prendre (α)



Conclusion

$p - value = 3{,}3\% < \alpha = 5\%$

On rejette H_0

5.5 – Tests d'adéquation du χ^2

Objectif

Savoir évaluer si une distribution observée s'ajuste à (ou : est en adéquation avec) une distribution ou loi théorique

5.6 – Test d'ajustement de Kolmogorov - Smirnov