# Tests statistiques pour la biologie

Yan Pautrat
Université Paris-Sud

2018-2019

# Organisation

# Déroulement du module : les séances

- cinq cours au total
  - dont trois ou quatre (suivant les groupes) avant le premier TD !
  - des exercices d'autoévaluation seront proposés à la fin des cours
- six séances de travaux dirigés de trois heures chacune [1]
  - dont certaines en binôme (biologiste et mathématicien) appelées "biomaths"

pourquoi les "biomaths" ? pour reconnecter maths et applications en biologie

1. en fait, cinq de trois heures et une de deux heures

# Déroulement du module : les évaluations

- deux sessions “exercices puis devoir noté” de WIMS
- un partiel “de maths”
- un devoir sur table “de biomaths”
- un examen “de maths et biomaths”

# Sessions de WIMS

deux sessions de WIMS :

1. révision de probabilités + début des tests
   un exercice ouvert du 14/09 à midi au 21/09 à midi
   un exercice ouvert du 26/09 à midi au 03/10 à midi
2. suite des tests statistiques
   ouvert du 09/11 à midi au 21/11 à midi

séance encadrée le 18/09 de 12h15 à 13h45 : s'inscrire auprès de L. Peillex

posez vos questions sur le forum !

# Déroulement du module : les notes

$$\text{note } 1^\circ \text{ session} = 0{,}55 \times \text{examen} + 0{,}25 \times \text{partiel} + 0{,}20 \times \text{cont. continu}$$

$$\text{note } 2^\circ \text{ session} = \text{examen } 2^\circ \text{ session}$$

$$\text{note contrôle continu} = \text{moyenne pondérée de WIMS 1, WIMS 2 et DST}$$

# Documents en ligne

- ces transparents (au fur et à mesure de l'avancement du cours)
- toutes ces informations dans les "modalités"
- les conseils d'utilisation de WIMS
- les énoncés d'exercices

**Attention :** dans les modalités, le plan de cours n'est pas à jour !

# Introduction

# Objectifs du cours

Les

**tests statistiques pour la biologie**

C'est-à-dire la méthode pour prendre des décisions sur la base d'observations expérimentales.

La difficulté est que lors de ces observations :

- ce que l'on observe dépend aussi de facteurs non contrôlés,
- il y a des erreurs expérimentales, également incontrôlées.

Il va falloir intégrer ces deux aspects à notre procédure de décision.

# Exemples de décisions à prendre

1. À des températures normales, chaque œuf d'alligator d'Amérique a autant de chances de donner un mâle qu'une femelle. D'un échantillon de 20 œufs exposés pendant l'incubation à une température inhabituellement élevée, naissent 14 mâles et 6 femelles. Doit-on en conclure que la hausse de température favorise la naissance de mâles ?
2. Vous arrivez en Norvège, et les dix premiers hommes (jeunes adultes) que vous rencontrez mesurent 180, 185, 184, 188, 182, 187, 182, 185, 181, 186 cm. Sachant que la taille moyenne des Français de la même catégorie est de 179 cm, doit-on en conclure que les Norvégiens sont en moyenne plus grands que les Français ?
3. On examine 1000 Français originaires du Pays basque, et on constate que 740 ont pour groupe sanguin $O$, 250 un groupe sanguin $A$, 5 un groupe sanguin $B$ et 5 un groupe sanguin $AB$. Sachant que la répartition dans l'ensemble de la population française est 43%, 45%, 9%, 3%, doit-on en conclure que la population basque présente une particularité dans la population française ?

# Les observations sont aléatoires :

1. quand on compte le nombre de mâles sur 20 œufs d'alligator d'Amérique,
   pour chaque œuf, mâle/femelle dépend d'une fécondation non observée, on va donc considérer que cela se fait au hasard
2. quand on calcule la moyenne des tailles de dix Norvégiens choisis au hasard,
   un Norvégien a une taille donnée mais quand on le choisit au hasard, il peut être plus ou moins grand, donc la moyenne des tailles des dix est aléatoire
3. quand on étudie la répartition des groupes sanguins de 1000 Français originaires du Pays basque,
   ici aussi on choisit chaque individu au hasard, donc le résultat final est aléatoire

# Conclusion ?

Même si on a l'impression que la réponse à tous ces "Peut-on en conclure. . . " est "oui", on pourrait très bien avoir fait des observations "atypiques" :

1. par hasard, avoir observé "beaucoup" de mâles,
2. par hasard, être tombé sur les dix Norvégiens les plus grands,
3. par hasard, être tombé sur "beaucoup " de Basques de groupe O.

Il va donc falloir quantifier ces hasards, ou en termes techniques discuter la significativité des observations.

# Plan du cours

La structure du cours sera :

1. principe général des tests (sans maths)
2. tests sur le paramètre $p$ d'une loi binomiale
3. tests sur la moyenne d'une loi normale (quand la variance est connue)
4. tests sur la variance et tests sur la moyenne (quand la variance est inconnue) d'une loi normale
5. tests du $\chi^2$ d'adéquation et d'indépendance.

On verra que parmi les exemples ci-dessus

l'exemple 1 entre dans le cadre du chapitre 2,

l'exemple 2 entre dans le cadre des chapitres 3 *ou* 4,

l'exemple 3 entre dans le cadre du chapitre 5.

# Et la technique ?

- les outils “techniques” sont la théorie des probabilités (lois binomiales, de Poisson, normales. . . ) vue en Terminale et/ou en UE 191
- en amphi il n’y aura que des rappels et les TDs commencent très tard, donc révisez par vous-même !
- quelques outils :
  - classe WIMS 1,
  - feuille de TD n°1
  - livres disponibles à la bibliothèque universitaire :
    - *Biostatistique* de Valleron (cote 519.077),
    - *Mathématiques et statistiques pour les sciences de la nature* de Biau, Droniou et Herzlich (cote 510.077),
    - *Statistique appliquée aux sciences de la vie* de Rousseau (cote 519.077).

# Principe général des tests (sans maths)

# Les ingrédients principaux

Nous allons détailler un test sur un exemple un peu intuitif, en nous arrêtant avant le point où il faudrait faire des maths.

Nous allons en particulier discuter trois des ingrédients principaux d'un test :

- les deux hypothèses possibles (soit c'est comme ci, soit c'est comme ça) entre lesquelles vous voulez discriminer,
- la variable de test, autrement dit la quantité obtenue par les observations et qui va servir à prendre la décision,
- la différence de comportement de la variable de test suivant que c'est une hypothèse ou l'autre qui est vraie.

Ces trois ingrédients vont directement mener à la *règle de décision*.

# Un exemple (un peu) intuitif

Vous jouez à pile-ou-face avec un inconnu. Après 50 parties, il a gagné 49 fois. Croyez-vous qu'il triche ?

# Raisonnement

Pourquoi pensez-vous qu'il triche ?

Sans doute parce que vous pensez que "49, c'est trop". Pourtant, un score de 49 ou plus, cela *peut* arriver sans tricher ! mais vous me direz qu'il y a *très peu de chances* que ça arrive.

En résumé, votre raisonnement tient donc en une phrase :

> si l'inconnu ne trichait pas, un score de 49 ou plus aurait *très peu de chances* de se produire.

# Raisonnement

Si l'on détaille un peu plus, vous vous êtes sans doute dit que :

- il y avait deux hypothèses possibles : soit il ne triche pas, soit il s'arrange pour avoir à chaque tour plus d'une chance sur deux de gagner ;
- vous alliez baser votre décision sur la seule observation disponible : le nombre de parties gagnées par l'inconnu ;
- ce nombre sera plus grand si l'inconnu triche que s'il ne triche pas.

Et vous avez décidé d'affirmer que l'individu triche parce que la probabilité qu'il aurait *sans tricher* de gagner un nombre de fois *supérieur ou égal* à 49 est trop faible.

vous avez là les hypothèses/la variable de test/le comportement et la règle de décision

# Début de formalisation

En formalisant à peine plus ce que nous avons déjà écrit :

- les deux hypothèses sont,
  - la première, qu'il a, à chaque partie, une chance sur deux de gagner,
  - la seconde, qu'il a, à chaque partie plus d'une chance sur deux de gagner ;
- la variable de test est le nombre de parties gagnées par l'inconnu sur les cinquante parties jouées ;
- ce nombre a tendance à être plus grand lorsque la deuxième hypothèse est vraie que lorsque c'est la première qui est vraie.

Et on décide de rejeter l'hypothèse qu'il ne triche pas parce que la probabilité d'un score $\geq 49$ *calculée en faisant l'hypothèse qu'il ne triche pas* est "très faible". [2]

2. (et en effet, on peut calculer que cette probabilité vaut $\simeq 4 \times 10^{-14}$)

# Formalisation des autres exemples

1. À des températures normales, chaque œuf d'alligator d'Amérique a autant de chances de donner un mâle qu'une femelle. D'un échantillon de 20 œufs exposés pendant l'incubation à une température inhabituellement élevée, naissent 14 mâles et 6 femelles. Doit-on en conclure que la hausse de température favorise la naissance de mâles ?
2. Vous arrivez en Norvège, et les dix premiers hommes (jeunes adultes) que vous rencontrez mesurent 180, 185, 184, 188, 182, 187, 182, 185, 181, 186 cm. Sachant que la taille moyenne des Français de la même catégorie est de 179 cm, doit-on en conclure que les Norvégiens sont en moyenne plus grands que les Français ?
3. On examine 1000 Français originaires du Pays basque, et on constate que 740 ont pour groupe sanguin $O$, 250 un groupe sanguin $A$, 5 un groupe sanguin $B$ et 5 un groupe sanguin $AB$. Sachant que la répartition dans l'ensemble de la population française est 43%, 45%, 9%, 3%, doit-on en conclure que la population basque présente une particularité dans la population française ?

# Formalisation des autres exemples : exemple 1

- première hypothèse : à température plus élevée, chaque œuf a encore autant de chances de donner un mâle qu'une femelle,
  seconde hypothèse : à température plus élevée, chaque œuf a plus de chances de donner un mâle qu'une femelle,
- variable de test : le nombre de mâles qui naissent des 20 œufs observés,
- différence de comportement : ce nombre a tendance à être plus grand lorsque la deuxième hypothèse est vraie que lorsque c'est la première hypothèse qui est vraie.

On va donc rejeter la première hypothèse si le nombre de mâles observé est "trop grand" pour être vraisemblable sous cette hypothèse.

# Formalisation des autres exemples

1. À des températures normales, chaque œuf d'alligator d'Amérique a autant de chances de donner un mâle qu'une femelle. D'un échantillon de 20 œufs exposés pendant l'incubation à une température inhabituellement élevée, naissent 14 mâles et 6 femelles. Doit-on en conclure que la hausse de température favorise la naissance de mâles ?
2. Vous arrivez en Norvège, et les dix premiers hommes (jeunes adultes) que vous rencontrez mesurent 180, 185, 184, 188, 182, 187, 182, 185, 181, 186 cm. Sachant que la taille moyenne des Français de la même catégorie est de 179 cm, doit-on en conclure que les Norvégiens sont en moyenne plus grands que les Français ?
3. On examine 1000 Français originaires du Pays basque, et on constate que 740 ont pour groupe sanguin $O$, 250 un groupe sanguin $A$, 5 un groupe sanguin $B$ et 5 un groupe sanguin $AB$. Sachant que la répartition dans l'ensemble de la population française est 43%, 45%, 9%, 3%, doit-on en conclure que la population basque présente une particularité dans la population française ?

# Formalisation des autres exemples : exemple 2

- première hypothèse : la taille moyenne des Norvégiens est la même que celle des Français,
  seconde hypothèse : la taille moyenne des Norvégiens est supérieure à celle des Français,
- variable de test : la moyenne des tailles des dix Norvégiens mesurés,
- différence de comportement : cette quantité a tendance à être plus grande lorsque la seconde hypothèse est vraie que lorsque c'est la première qui est vraie.

On va donc rejeter la première hypothèse si la taille moyenne des dix Norvégiens observés est "trop élevée" pour être vraisemblable sous cette hypothèse.

# Formalisation des autres exemples

1. À des températures normales, chaque œuf d'alligator d'Amérique a autant de chances de donner un mâle qu'une femelle. D'un échantillon de 20 œufs exposés pendant l'incubation à une température inhabituellement élevée, naissent 14 mâles et 6 femelles. Doit-on en conclure que la hausse de température favorise la naissance de mâles ?
2. Vous arrivez en Norvège, et les dix premiers hommes (jeunes adultes) que vous rencontrez mesurent 180, 185, 184, 188, 182, 187, 182, 185, 181, 186 cm. Sachant que la taille moyenne des Français de la même catégorie est de 179 cm, doit-on en conclure que les Norvégiens sont en moyenne plus grands que les Français ?
3. On examine 1000 Français originaires du Pays basque, et on constate que 740 ont pour groupe sanguin $O$, 250 un groupe sanguin $A$, 5 un groupe sanguin $B$ et 5 un groupe sanguin $AB$. Sachant que la répartition dans l'ensemble de la population française est 43%, 45%, 9%, 3%, doit-on en conclure que la population basque présente une particularité dans la population française ?

# Formalisation des autres exemples : exemple 3

- première hypothèse : la répartition des groupes sanguins est la même chez les Basques que dans l'ensemble des Français,
  seconde hypothèse : la répartition des groupes sanguins est différente chez les Basques que dans l'ensemble des Français,
- variable de test : ? ? ?
  on sent bien qu'on va utiliser les écarts entre les proportions observées et les proportions parmi l'ensemble des Français, mais comment ?
- différence de comportement : ? ? ?

La variable de test à utiliser sera donnée par un théorème mathématique au chapitre 5.

# Hypothèses $H_0/H_1$

Il est habituel d'appeler $H_0$ et $H_1$ les deux hypothèses.

Les hypothèses $H_0$ et $H_1$ *ne sont pas interchangeables*
(et on verra pourquoi plus tard).

Comment choisir alors qui est $H_0$ et qui est $H_1$ ?

# Hypothèses $H_0/H_1$

Si on a déjà les deux hypothèses il y a deux contraintes pour choisir laquelle est $H_0$ et laquelle est $H_1$ :

1. $H_0$ doit être l'hypothèse à priori, et $H_1$ l'hypothèse qu'on n'accepte que si l'on a de bonnes raisons de le faire.
   à priori, température et sexe des bébés ne sont pas liés, et Norvégiens ou Basques ne sont pas différents des Français ; pour affirmer le contraire, il faudrait de bons arguments
2. $H_0$ doit donner les valeurs des paramètres (on expliquera plus tard pourquoi),
   autrement dit, il peut y avoir des inégalités sur les paramètres dans $H_1$ mais pas dans $H_0$

(il peut arriver que ces deux contraintes ne soient pas tout à fait compatibles)

la première contrainte fait que l'on parle pour la conclusion de *conserver* $H_0$ ou de *rejeter* $H_0$ mais pas d'accepter $H_1$

# Hypothèses $H_0/H_1$

Ces contraintes imposent en général ce qu'est $H_0$ ; le choix de $H_1$ est plus une question de biologie que de mathématiques !

Si par exemple on veut tester l'effet de l'alcool sur une certaine capacité, et que l'on fait passer pour cela une épreuve à des buveurs, l'hypothèse $H_0$ sera en général

$H_0$ : un buveur a les mêmes chances de succès qu'un non buveur

mais $H_1$ dépendra sans doute de la nature de la capacité !

- si l'on teste une capacité de raisonnement ou de coordination, on choisira

  $H_1$ : un buveur a des chances de succès moindres qu'un non buveur

- si l'on teste une capacité à se déshabiller en public, on choisira

  $H_1$ : un buveur a des chances de succès supérieures à un non buveur

# Hypothèses $H_0/H_1$

Dans l'exemple 2, on a décidé de tester

$H_0$ : la taille moyenne des Norvégiens est la même que celle des Français,

contre

$H_1$ : la taille moyenne des Norvégiens est supérieure à celle des Français,

en particulier parce qu'un préjugé nous dit que *si* les Norvégiens sont différents des Français, alors ils sont plus grands.

# Hypothèses $H_0/H_1$

Si par exemple on avait mesuré dix ~~Norvégiens~~ Boliviens qui faisaient 160, 165, 164, 168, 162, 167, 162, 165, 161, 166 cm, on aurait pu décider de tester

$H_0$ : la taille moyenne des Boliviens est la même que celle des Français,

contre

$H_1'$ : la taille moyenne des Boliviens est inférieure à celle des Français.

Auquel cas on aurait dit que sous $H_1'$ la variable de test[3] a tendance à être plus petite que sous $H_0$, et on aurait rejeté $H_0$ si l'on avait observé des valeurs "trop petites".

3. qui est toujours la moyenne des tailles des dix individus observés

# Hypothèses $H_0/H_1$ : tests unilatères vs. bilatères

Si l'on n'avait pas d'indication à priori sur la taille des ~~Norvégiens~~ Italiens, on pourrait vouloir tester

$H_0$ : la taille moyenne des Italiens est la même que celle des Français,

contre

$H_1''$ : la taille moyenne des Italiens est différente à celle des Français.

On aurait alors dit que sous $H_1'$ la variable de test [4] a tendance à être plus petite ou plus grande que sous $H_0$, et on aurait rejeté $H_0$ si l'on avait observé des valeurs "trop petites ou trop grandes".

(c'est un peu plus compliqué et nous reviendrons là-dessus plus tard).

Un tel test est appelé *bilatère*, les tests précédents sont appelés *unilatères*.

4. qui est toujours la moyenne des tailles des dix individus observés

# Probabilité “petite” : le niveau

On a dit ci-dessus que l’on “rejetait l’hypothèse $H_0$” si (en gros) le comportement observé aurait, si $H_0$ était vraie, une probabilité “trop petite”.

Mais que signifie “trop petite” ?

En général on décidera à l’avance ce que l’on appelle une probabilité “petite”. La valeur choisie est appelée le *niveau* du test (que l’on note en général $\alpha$).

On prendra souvent un niveau de 5% mais c’est très contestable !

# Pourquoi un niveau à priori ?

Vous jouez à pile-ou-face avec un inconnu. Après 50 parties, il a gagné 49 fois/42 fois/38 fois/31 fois. Croyez-vous qu'il triche ?

Se donner un niveau permet de déterminer (avant observation) à partir de quel score on accusera l'autre de tricher. Cette valeur est appelée le *seuil*.

# Pourquoi un niveau à priori ?

Les mathématiciens ont l'habitude de travailler en calculant le seuil car cela permet de déterminer la règle de décision *avant* d'avoir les observations.

Dans l'exemple des cinquante parties de pile-ou-face, on peut calculer que :

- la probabilité d'un score $\geq 30$ avec une pièce équilibrée est $\simeq 10{,}1\%$.
- la probabilité d'un score $\geq 31$ avec une pièce équilibrée est $\simeq 5{,}9\%$.
- la probabilité d'un score $\geq 32$ avec une pièce équilibrée est $\simeq 3{,}2\%$.
- la probabilité d'un score $\geq 33$ avec une pièce équilibrée est $\simeq 1{,}6\%$.

Si l'on a choisi un niveau $\alpha = 5\%$ (autrement dit, si l'on considère "improbable" tout événement de proba $\leq 5\%$), alors la valeur à partir de laquelle on trouvera louche le score de l'inconnu est 32.

# Influence du choix du seuil

D'après les calculs cités ci-dessus, dans l'exemple du pile-ou-face :

au niveau $\alpha = 7\%$ on accuse l'inconnu de tricher s'il fait 31 ou plus,

au niveau $\alpha = 5\%$ on accuse l'inconnu de tricher s'il fait 32 ou plus,

au niveau $\alpha = 2\%$ on accuse l'inconnu de tricher s'il fait 33 ou plus.

De manière plus générale :

plus le niveau est petit et plus on a tendance à conserver $H_0$

(car "il faut plus de preuves" pour rejeter $H_0$).

# Et maintenant ?

Vous avez déjà le principe général de tous les tests que nous allons étudier. Il vous manque la technique permettant de trouver le seuil d'un test. C'est ce que nous allons voir à partir de maintenant.

# Chapitre 1 : exercices d'autoévaluation

Quelles sont les hypothèses $H_0$ et $H_1$ dans les pages 22, 24 et 26 ?

Réponse : à chaque fois $H_0$ est la première hypothèse et $H_1$ la deuxième.

# Chapitre 1 : exercices d'autoévaluation

Donner hypothèses, variable de test, comportement de cette variable sous $H_0/H_1$ et "forme" de la règle de décision dans les trois cas suivants :

1. On s'intéresse à la présence à Paris d'une maladie rare. On sait qu'en temps normal, chaque Parisien a une probabilité $2 \times 10^{-6}$, de contracter cette maladie. Cette année, on a observé sept nouveaux cas. Peut-on en conclure que la maladie est devenue plus virulente ?
2. On veut tester si le tabac diminue les performances sportives. Pour cela, on fait passer à vingt fumeurs une série d'épreuves et on leur attribue en fin de compte une note sur 500. L'épreuve est normalisée de telle manière que la note moyenne de l'ensemble de la population est de 250. Les vingt fumeurs obtiennent une note de 236. Que peut-on en conclure ?
3. On veut tester l'influence de l'alcool sur une certaine capacité. Pour cela, on réunit 30 individus, on les fait boire puis leur fait effectuer un test. Ce test est réussi par un individu à jeun dans 40% des cas. Ici, 7 individus réussissent.

# Chapitre 1 : exercices d'autoévaluation

Réponses :

1. on prend $H_0$ : la maladie n'est pas plus virulente que d'habitude contre $H_1$ : la maladie est plus virulente que d'habitude ; la variable de test est le nombre de malades cette année ; cette variable prend des valeurs plus grandes si $H_1$ est vraie que si c'est $H_0$ qui est vraie.
2. on prend $H_0$ : le tabac n'a pas d'effet sur les performances sportives contre $H_1$ : le tabac a un effet ; la variable de test est le score moyen obtenu par les vingt fumeurs ; cette variable prend des valeurs plus petites si $H_1$ est vraie que si c'est $H_0$ qui est vraie.
3. on prend $H_0$ : l'alcool n'a pas d'effet sur cette capacité. Pour $H_1$, ça dépend ! (voir page 29). Si par exemple on teste une capacité de raisonnement, on prendra $H_1$ : l'alcool diminue cette capacité. La variable de test est le nombre d'individus qui réussissent le test. *Avec le choix de* $H_0/H_1$ *ci-dessus*, cette variable prend des valeurs plus petites si $H_1$ est vraie que si c'est $H_0$ qui est vraie. Si l'on avait testé la capacité à se déshabiller en public, on aurait pris $H_1$ : l'alcool augmente cette capacité et la variable de test aurait pris des valeurs plus grandes lorsque $H_1$ est vraie que lorsque c'est $H_0$ qui est vraie

# Chapitre 1 : exercices d'autoévaluation

Supposons que dans les exercices ci-dessus :

1. Au niveau 10%, on conserve l'hypothèse que la maladie n'est pas plus virulente cette année que d'habitude. Que conclurait-on au niveau 5% ?
2. Au niveau 3%, on rejette l'hypothèse que le tabac n'a pas d'influence sur les capacités sportives. Que peut-on dire au niveau 5% ?

Réponse : les conclusions ne changent pas car si l'on conserve $H_0$ à un certain niveau, on la conserve d'autant plus avec un niveau inférieur et inversement, si l'on rejette $H_0$ à un certain niveau, on la rejette d'autant plus avec un niveau supérieur. Voir page 37.

# Tests sur le paramètre $p$ d'une loi binomiale

# Pour quoi faire ?

Les résultats de cette section permettent de traiter l'exemple numéro 1 :

À des températures normales, chaque œuf d'alligator d'Amérique a autant de chances de donner un mâle qu'une femelle. D'un échantillon de 20 œufs exposés pendant l'incubation à une température inhabituellement élevée, naissent 14 mâles et 6 femelles. Doit-on en conclure que la hausse de température favorise la naissance de mâles ?

# Retour sur l'exemple du pile-ou-face

Vous jouez à pile-ou-face avec un inconnu. Après 50 parties, il a gagné 49 fois/42 fois/38 fois/31/g fois. Croyez-vous qu'il triche ?

Pour traiter précisément cet exemple, il nous faut revenir sur les *lois binomiales*.

# Rappels : loi binomiale

- une **expérience de Bernoulli de paramètre** $p$ est une expérience aléatoire à deux issues "succès" et "échec" dans laquelle la probabilité de succès est $p$ (de valeur entre 0 et 1) et la probabilité d'échec $1-p$
- si l'on répète $n$ fois et de manière indépendante une expérience de Bernoulli de paramètre $p$, le nombre total $N$ de succès est aléatoire et suit une **loi binomiale de paramètres** $n$ **et** $p$, notée $\mathcal{B}(n;p)$
- si $N \sim \mathcal{B}(n;p)$ alors $N$ peut prendre toutes les valeurs $0,1,\ldots,n$ et

$$\mathbb{P}(N=k) = \binom{n}{k} p^k (1-p)^{n-k} \text{ pour } k \text{ dans } 0,1,\ldots,n.$$

# Rappels : loi binomiale

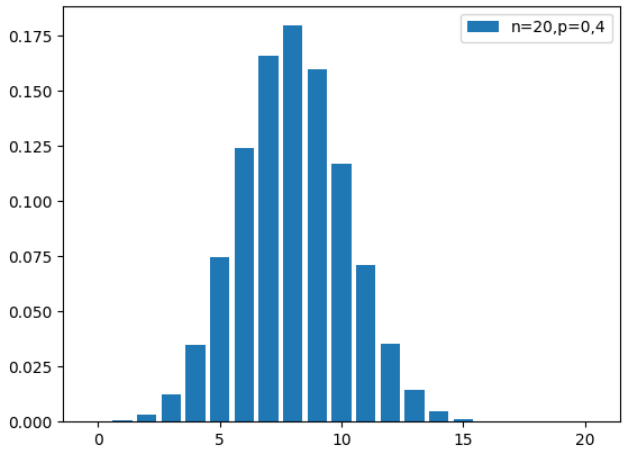
**Exemples :**

- on joue cinquante fois à pile ou face ; si à chaque lancer la probabilité de faire "pile" est ~~0,5~~ $p$ alors le nombre total de "pile" sur les cinquante lancers est aléatoire et suit une loi binomiale ~~$\mathcal{B}(50; 0{,}5)$~~ $\mathcal{B}(50; p)$
- on observe l'éclosion de vingt œufs d'alligator d'Amérique ; si chaque œuf a une probabilité ~~0,5~~ $p$ de donner un mâle, le nombre total de mâles qui sort des vingt œufs est aléatoire et suit une loi binomiale ~~$\mathcal{B}(20; 0{,}5)$~~ $\mathcal{B}(20; p)$
- on fait passer une épreuve à trente buveurs ; si chacun a une probabilité ~~0,4~~ $p$ de réussir l'épreuve, alors le nombre de buveurs qui la réussit est aléatoire et suit une loi binomiale ~~$\mathcal{B}(30; 0{,}4)$~~ $\mathcal{B}(30; p)$.

Ici on suppose implicitement que les lancers/naissances/épreuves se font de manière *identique* et *indépendante*. N'oubliez pas ces hypothèses quand vous définissez vos plans d'expérience !

# Rappels : loi binomiale

allure de cette distribution (ici pour $n = 20$, $p = 0{,}4$) :




les hauteurs des bâtons donnent les probabilités d'*une* valeur :

$$\mathbb{P}(N = 5) \simeq 0{,}075 \quad \mathbb{P}(N = 6) \simeq 0{,}124 \quad \mathbb{P}(N = 7) \simeq 0{,}165$$

on peut alors retrouver la probabilité d'un intervalle :

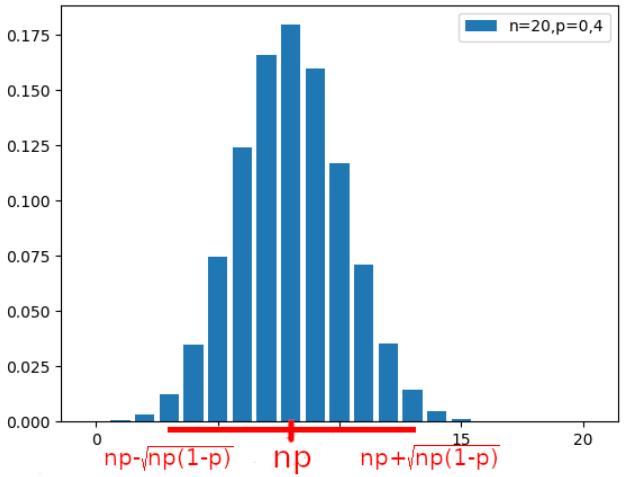
$$\mathbb{P}(5 \leq N \leq 7) = \mathbb{P}(N = 5) + \mathbb{P}(N = 6) + \mathbb{P}(N = 7) \simeq 0{,}364.$$

on calcule ces hauteurs soit avec calculatrice/ordinateur et la formule donnée plus tôt, soit avec les tables de valeurs numériques (voir plus loin)

# Rappels : loi binomiale, dépendance en $n$ et $p$

la binomiale $\mathcal{B}(n;p)$ est centrée en $np$

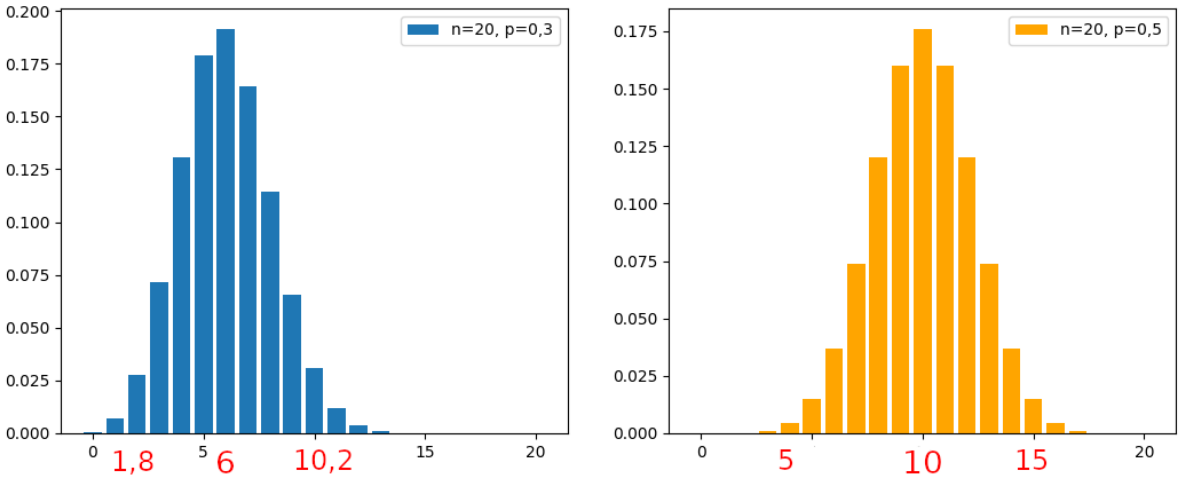
et prend "la majorité" de ses valeurs sur $[np-\sqrt{np(1-p)}, np+\sqrt{np(1-p)}]$ :




(ici $np-\sqrt{np(1-p)} \simeq 3{,}2$ et $np+\sqrt{np(1-p)} \simeq 12{,}8$)

# Rappels : loi binomiale, dépendance en $n$ et $p$

en particulier à $n$ fixé, les valeurs probables sont plus élevées quand $p$ augmente :




(en rouge les valeurs $np$ et $np \pm \sqrt{np(1-p)}$)

C'est normal : si on fait le même nombre d'essais mais qu'on a plus de chances de réussir à chaque fois, on a tendance à réussir plus de fois !

# Rappels : loi binomiale, espérance et variance

pour une binomiale $N \sim \mathcal{B}(n; p)$ :

l'espérance, qui représente la "moyenne pondérée" des valeurs possibles, est

$$\mathbb{E}(N) \overset{\text{déf}}{=} \sum_{k=0}^{n} k\,\mathbb{P}(N = k) \quad \text{vaut} \quad \mathbb{E}(N) = n \times p$$

la variance, qui donne une information sur la "dispersion" des valeurs possibles, est

$$\mathbb{V}(N) \overset{\text{déf}}{=} \sum_{k=0}^{n} \big(k - \mathbb{E}(N)\big)^2\,\mathbb{P}(N = k) \quad \text{vaut} \quad \mathbb{V}(N) = n \times p \times (1 - p)$$

# Exemple du pile-ou-face, suite

Si l'on note $p$ la probabilité qu'a l'inconnu de gagner *une* partie, alors son nombre total $N$ de victoires suit une loi binomiale $\mathcal{B}(50,p)$.

La formalisation donnée page 20 devient

- les deux hypothèses sont,
  - $H_0 : p = 0{,}5$
  - $H_1 : p > 0{,}5$
- la variable de test est le nombre $N$ ;
- ce nombre a tendance a être plus grand lorsque $H_1$ est vraie que lorsque c'est $H_0$ qui est vraie.

On a déjà dit que l'on décidait de rejeter $H_0$ si l'on observait des valeurs de $N$ "trop grandes".

# Exemple du pile-ou-face, suite

Que veut dire "trop grand" ? comme on l'a dit pages 33 et 36, on a l'habitude de travailler en prenant un *niveau* et en déterminant le *seuil*.

Si l'on note $\alpha$ le niveau, alors ici le seuil est le plus petit $n_0$ tel que la probabilité de faire un score supérieur ou égal à $n_0$ si c'est $H_0$ qui est vraie soit inférieure ou égale à $\alpha$.

On note

$$\mathbb{P}_{H_0}(N \geq n_0) \quad \overset{\text{notation}}{=} \quad \text{la probabilité de faire un score supérieur ou égal à } n_0 \text{ si c'est } H_0 \text{ qui est vraie}$$

Le seuil est donc le plus petit $n_0$ tel que $\mathbb{P}_{H_0}(N \geq n_0) \leq \alpha$.

Comment calculer ce seuil ?

# Exemple du pile-ou-face, suite

Quand $H_0$ est vraie, la variable $N$ suit une loi binomiale $\mathcal{B}(50; 0{,}5)$.

On est donc en train de chercher le plus petit $n_0$ pour lequel $\mathbb{P}(N \geq n_0) \leq \alpha$ lorsque $N \sim \mathcal{B}(50; 0{,}5)$.

Ce $n_0$ peut s'obtenir à partir des *tables de valeurs numériques*... même s'il est un peu caché.

# Rappels : tables numériques et règles de calcul

Les tables de valeurs numériques donnent (en général) les valeurs de la *fonction de répartition* $k \mapsto \mathbb{P}(X \leq k)$ pour les différentes valeurs de $k$, en fonction de la loi de $X$.

Par conséquent, il est facile de :

- calculer une probabilité $\mathbb{P}(X \leq k)$ pour $X \sim \mathcal{B}(n; p)$ : on lit directement la valeur. Par exemple, pour $X \sim \mathcal{B}(10; 0{,}25)$ on a $\mathbb{P}(X \leq 3) = 0{,}7759$

| $n = 10$ | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $k \setminus p$ | 0.005 | 0.01 | 0.05 | 0.10 | 0.15 | 0.20 | 0.25 | 0.30 | 0.35 | 0.40 | 0.45 |
| 0 | 0.9511 | 0.9044 | 0.5987 | 0.3487 | 0.1969 | 0.1074 | 0.0563 | 0.0282 | 0.0135 | 0.0060 | 0.0025 |
| 1 | 0.9989 | 0.9957 | 0.9139 | 0.7361 | 0.5443 | 0.3758 | 0.2440 | 0.1493 | 0.0860 | 0.0464 | 0.0233 |
| 2 | 1.0000 | 0.9999 | 0.9885 | 0.9298 | 0.8202 | 0.6778 | 0.5256 | 0.3828 | 0.2616 | 0.1673 | 0.0996 |
| 3 | 1.0000 | 1.0000 | 0.9990 | 0.9872 | 0.9500 | 0.8791 | 0.7759 | 0.6496 | 0.5138 | 0.3823 | 0.2660 |
| 4 | 1.0000 | 1.0000 | 0.9999 | 0.9984 | 0.9901 | 0.9672 | 0.9219 | 0.8497 | 0.7515 | 0.6331 | 0.5044 |

- chercher le plus grand $k$ tel que $\mathbb{P}(X \leq k) \leq \alpha$ : on lit les valeurs successives de $\mathbb{P}(X \leq k)$ dans le tableau

le plus grand $k$ tel que $\mathbb{P}(X \leq k) \leq 0{,}05$ pour $X \sim \mathcal{B}(20; 0{,}3)$ est $k = 2$ car

$$\mathbb{P}(X \leq 2) = 0{,}0355 \leq 0{,}05$$
$$\mathbb{P}(X \leq 3) = 0{,}1071 > 0{,}05$$

| $k \setminus p$ | $n = 20$ | | | |
|---|---|---|---|---|
| | 0.15 | 0.20 | 0.25 | 0.30 |
| 0 | 0.0388 | 0.0115 | 0.0032 | 0.0008 |
| 1 | 0.1756 | 0.0692 | 0.0243 | 0.0076 |
| 2 | 0.4049 | 0.2061 | 0.0913 | 0.0355 |
| 3 | 0.6477 | 0.4114 | 0.2252 | 0.1071 |
| 4 | 0.8298 | 0.6296 | 0.4148 | 0.2375 |
| 5 | 0.9327 | 0.8042 | 0.6172 | 0.4164 |
| 6 | 0.9781 | 0.9133 | 0.7858 | 0.6080 |
| 7 | 0.9941 | 0.9679 | 0.8982 | 0.7723 |

# Rappels : tables numériques et règles de calcul

Si c'est autre chose que l'on cherche, on applique les règles suivantes :

- on peut obtenir les valeurs de $\mathbb{P}(X \geq k)$ en utilisant la relation

$$\mathbb{P}(X \geq k) = 1 - \mathbb{P}(X < k) = 1 - \mathbb{P}(X \leq k-1)$$

- on peut obtenir les valeurs de $\mathbb{P}(X = k)$ en utilisant la relation

$$\mathbb{P}(X = k) = \mathbb{P}(X \leq k) - \mathbb{P}(X \leq k-1)$$

(ces règles sont valables pour les variables aléatoires *à valeurs entières*)

# Exemple du pile-ou-face, suite

On cherche lorsque $N \sim \mathcal{B}(50; 0{,}5)$ le plus petit $n_0$ pour lequel

$$\mathbb{P}(N \geq n_0) \leq \alpha.$$

On utilise le fait que

$$\mathbb{P}(N \geq n_0) = 1 - \mathbb{P}(N < n_0) = 1 - \mathbb{P}(N \leq n_0 - 1)$$

et donc

$$\mathbb{P}(N \geq n_0) \leq \alpha \Leftrightarrow \mathbb{P}(N \leq n_0 - 1) \geq 1 - \alpha.$$

On cherche donc lorsque $N \sim \mathcal{B}(50; 0{,}5)$ le plus petit $n_0$ pour lequel

$$\mathbb{P}(N \leq n_0 - 1) \geq 1 - \alpha$$

et ça on sait faire !

# Exemple du pile-ou-face, suite

par exemple en lisant dans le tableau $n = 50$, colonne $p = 0{,}5$ :

- au niveau $\alpha = 10\%$ on trouve $n_0 - 1 = 30$ donc $n_0 = 31$,
- au niveau $\alpha = 2\%$ on trouve $n_0 - 1 = 32$ donc $n_0 = 33$,
- au niveau $\alpha = 0{,}5\%$ on trouve $n_0 - 1 = 34$ donc $n_0 = 35$.

| | $n = 50$ | |
|---|---|---|
| $k \setminus p$ | 0.45 | 0.5 |
| 28 | 0.9556 | 0.8389 |
| 29 | 0.9765 | 0.8987 |
| 30 | 0.9884 | 0.9405 |
| 31 | 0.9947 | 0.9675 |
| 32 | 0.9978 | 0.9836 |
| 33 | 0.9991 | 0.9923 |
| 34 | 0.9997 | 0.9967 |
| 35 | 0.9999 | 0.9987 |
| 36 | 1.0000 | 0.9995 |

On va donc rejeter l'hypothèse $H_0$ : l'inconnu ne triche pas

- au niveau 10%, s'il fait un score $\geq 31$,
- au niveau 2%, s'il fait un score $\geq 33$,
- au niveau 0,5%, s'il fait un score $\geq 35$.

On sait donc faire des *tests sur le paramètre p d'une loi binomiale*. On va maintenant les réécrire "proprement".

# Un dernier mot sur les tables numériques

Il peut arriver que l'on n'aie pas la table de la binomiale $\mathcal{B}(n; p)$ pour les valeurs des paramètres $n$ et $p$ qui nous intéressent :

- si c'est parce que $p > 0{,}5$ ; il faut alors utiliser le fait que $n - X \sim \mathcal{B}(n; 1-p)$ et que, par exemple, $\mathbb{P}(X \geq n_0) = \mathbb{P}(n - X \leq n - n_0)$.
- Si c'est parce que $n$ ou $p$ ne sont pas *exactement* dans le tableau, on prend des valeurs proches (par exemple pour $n = 49$ et $p = 0{,}51$ on regardera le tableau pour $n = 50$ et $p = 0{,}5$).
- Si c'est parce que $n$ est grand, alors on peut utiliser l'approximation de la loi binomiale par une loi normale (voir plus loin).

# Rédaction des tests en sept points

Nous vous demanderons toujours de rédiger un test suivant les sept points ci-dessous :

1. le modèle
2. les hypothèses
3. la variable de test
4. la forme de la zone de rejet
5. le seuil de la zone de rejet (suivant le niveau)
6. la conclusion (suivant le niveau et les observations)
7. la $p$-valeur (suivant les observations).

Plutôt que de définir chacun de ces termes, voyons ce qu'ils signifient sur notre exemple.

# Rédaction des tests en sept points

Pour les points 1 à 4 le chapitre 1 suffit ! et vous n'avez pas besoin d'avoir ni le niveau du test, ni les observations.

Pour le point 5 il faut faire un peu de maths, et il vous faut le niveau du test.

Pour les points 6 et 7 il vous faut les résultats d'observation.

# Rédaction en sept points : notre exemple

(supposons que l'inconnu ait gagné 38 fois, et faisons le test au niveau $\alpha = 2\%$)

1. **le modèle** : l'inconnu répète 50 fois et de manière indépendante une expérience de Bernoulli de paramètre $p$ (représentant la probabilité qu'a l'inconnu de gagner). On note $N$ le nombre de succès.
2. **les hypothèses** : $H_0$ : $p = 0{,}5$ et $H_1$ : $p > 0{,}5$.
3. **la variable de test** : on utilise la variable $N$.
4. **la forme de la zone de rejet** : comme $N$ est plutôt plus grand sous l'hypothèse $H_1$ que sous l'hypothèse $H_0$, on va rejeter $H_0$ si $N$ est "trop grand". La zone de rejet est donc de la forme $[N \geq n_0]$.
5. **le seuil de la zone de rejet** : au niveau $\alpha = 2\%$, on trouve (ce sont les calculs en pages 60 et 61) que le plus petit $n_0$ tel que $\mathbb{P}_{H_0}(N \geq n_0) \leq \alpha$ est $n_0 = 33$.
6. **la conclusion** : puisqu'on a observé $N = 38$, on rejette $H_0$ au niveau 2%.
7. **la $p$-valeur** : $\mathbb{P}_{H_0}(N \geq 38) \simeq 0{,}02\%$ donc la $p$-valeur vaut 0,02%.

# La $p$-valeur

La $p$-valeur est l'élément le plus mystérieux ou le plus simple, au choix.

- C'est le niveau pour lequel le seuil est égal à la valeur observée.
- En pratique : si le test est "on rejette si $N$ est trop grand" et qu'on a observé $N = 38$, c'est simplement la probabilité sous $H_0$ d'observer $N \geq 38$.
- Les tests "à la biologiste" reviennent à calculer la $p$-valeur puis à rejeter $H_0$ si elle est inférieure au niveau, à conserver $H_0$ si elle est supérieure.
- La $p$-valeur donne donc plus d'information que la conclusion : si elle est très proche du niveau, c'est que la conclusion a été choisie "de justesse" ; si elle est très différente, c'est que la conclusion est choisie "sans hésitation".

# Un autre exemple dans le cas discret

On cherche à estimer si l'alcool a un effet sur les capacités d'un individu. Pour cela, on réunit 30 individus, on les fait boire puis leur fait effectuer un test. Ce test est réussi par un individu à jeun dans 40% des cas. Ici, 7 individus réussissent. Que peut-on en conclure au niveau 5% ?

On va analyser cet exemple.

*Ce qui suit ne constitue pas une rédaction correcte mais simplement une analyse ! il faudrait rédiger le tout en sept points.*

# Exemple : effets de l'alcool

On reprend les différents points de l'analyse faite pour le pile-ou-face

- on note $p$ la probabilité de réussir le test pour un individu qui a bu de l'alcool. Dans ce cas, le nombre $R$ d'individus (sur 30) qui réussit le test suit une loi binomiale $\mathcal{B}(30; p)$
- on veut tester $H_0 : p = 0{,}4$ contre... quoi ? on peut considérer
  - que l'alcool est soit neutre, soit nocif pour la capacité testée. Dans ce cas les hypothèses sont $H_0 : p = 0{,}4$ contre $H_1 : p < 0{,}4$
  - que l'alcool est soit neutre, soit bénéfique pour la capacité testée. Dans ce cas les hypothèses sont $H_0 : p = 0{,}4$ contre $H_1 : p > 0{,}4$
  - que l'alcool est soit neutre, soit actif (dans un sens ou dans l'autre). Dans ce cas, les hypothèses sont $H_0 : p = 0{,}4$ contre $H_1 : p \neq 0{,}4$

  Encore une fois, le choix des hypothèses est une question de biologie, pas de mathématiques !

# Exemple : effets de l'alcool (neutre ou nocif)

*On poursuit cet exercice en nous plaçant dans le premier cas*. On choisit donc de tester $H_0 : p = 0{,}4$ contre $H_1 : p < 0{,}4$.

- $R$ est plutôt plus *petit* sous $H_1$ que sous $H_0$, on va donc rejeter $H_0$ si $R$ est trop petit, la zone de rejet sera de la forme $[R \leq r_0]$
- avec les tables on trouve que le plus *grand* $r_0$ tel que $\mathbb{P}_{H_0}(R \leq r_0) \leq 5\%$ est $r_0 = 7$
- on observe bien un $R \leq 7$, donc on rejette l'hypothèse $H_0$ au niveau 5%
- la $p$-valeur vaut $\mathbb{P}_{H_0}(R \leq 7)$ et d'après la table c'est $\simeq 4{,}3\%$.

# Exemple : effets de l'alcool (neutre ou bénéfique)

*On poursuit cet exercice en nous plaçant dans le deuxième cas*. On choisit donc de tester $H_0 : p = 0{,}4$ contre $H_1 : p > 0{,}4$.

- $R$ est plutôt plus *petit* sous $H_1$ que sous $H_0$, on va donc rejeter $H_0$ si $R$ est trop grand, la zone de rejet sera de la forme $[R \geq r_0]$
- on cherche le plus *petit* $r_0$ tel que $\mathbb{P}_{H_0}(R \geq r_0) \leq 5\%$. Or

$$\mathbb{P}_{H_0}(R \geq r_0) = 1 - \mathbb{P}_{H_0}(R < r_0) = 1 - \mathbb{P}_{H_0}(R \leq r_0 - 1)$$

donc

$$\mathbb{P}_{H_0}(R \geq r_0) \leq 5\% \Leftrightarrow \mathbb{P}_{H_0}(R \leq r_0 - 1) \geq 95\%$$

et la table nous dit que le plus petit $r_0$ tel que $\mathbb{P}_{H_0}(R \leq r_0 - 1) \geq 95\%$ vérifie $r_0 - 1 = 16$ donc $r_0 = 17$.

- on observe un $R < 17$, donc on ne rejette pas l'hypothèse $H_0$ au niveau 5%
- la $p$-valeur vaut $\mathbb{P}_{H_0}(R \geq 7) = 1 - \mathbb{P}_{H_0}(R < 7) = 1 - \mathbb{P}_{H_0}(R \leq 6)$ et d'après la table c'est $\simeq 1 - 0{,}0172 \simeq 98{,}3\%$. On ne doute pas de notre conclusion !

# Exemple : effets de l'alcool (neutre ou actif)

Et si l'on avait choisi comme hypothèses $H_0 : p = 0{,}4$ contre $H_1 : p \neq 0{,}4$ ?

- comment se comporte $R$ en fonction de $p$ ?
  - si $p < 0{,}4$ alors $R$ est plutôt plus *petit* sous $H_1$ que sous $H_0$,
  - mais si $p > 0{,}4$ alors $R$ est plutôt plus *grand* sous $H_1$ que sous $H_0$

# Exemple : effets de l'alcool (neutre ou actif)

- La zone de rejet va donc être de la forme $[R \leq r_1$ **ou** $R \geq r_2]$
- on cherche $r_1$ et $r_2$ tels que $\mathbb{P}_{H_0}(R \leq r_1 \text{ ou } R \geq r_2) \leq 5\%$. Il existe plusieurs solutions mais la plus naturelle consiste à choisir
  - $r_1$ le plus grand possible tel que $\mathbb{P}_{H_0}(R \leq r_1) \leq 2{,}5\%$
  - $r_2$ le plus petit possible tel que $\mathbb{P}_{H_0}(R \geq r_2) \leq 2{,}5\%$

  (où l'on choisit 2,5% parce que c'est la moitié de 5%). Avec les tables on trouve par le même genre de calculs qu'auparavant $r_1 = 6$ et $r_2 = 18$.
- on observe un $R$ qui n'est ni $\leq 6$, ni $\geq 18$ donc on ne rejette pas l'hypothèse $H_0$ au niveau 5%.

# Exemple : effets de l'alcool (neutre ou actif)

- le calcul de la $p$-valeur est également un peu plus compliqué que dans les cas précédents. Remarquez que les seuils $r_1$ et $r_2$ ont été déterminés en fonction du niveau $\alpha$ par

$$\mathbb{P}_{H_0}(R \leq r_1) \leq \alpha/2 \qquad \mathbb{P}_{H_0}(R \geq r_2) \leq \alpha/2$$

encore une fois on remplace dans ces expressions $r_1$ et $r_2$ par la valeur observée. Cela nous donne deux probabilités, on garde la plus petite des deux probabilités et cela donne la moitié de la $p$-valeur. Ici $\mathbb{P}_{H_0}(R \leq 7) \simeq 4{,}3\%$ et $\mathbb{P}_{H_0}(R \geq 7) = 1 - \mathbb{P}_{H_0}(R \leq 6)$ est beaucoup plus grand, donc la $p$-valeur vaut $2 \times 4{,}3\% = 8{,}6\%$.

# Hypothèses et conclusions ?

On remarque qu'avec le même niveau $\alpha = 5\%$ et la même observation $R = 7$ :

si l'on teste $H_0 : p = 0{,}4$ contre $H_1 : p < 0{,}4$ alors on rejette $H_0$,

si l'on teste $H_0 : p = 0{,}4$ contre $H_1 : p \neq 0{,}4$ alors on ne rejette pas $H_0$

alors que $H_0$ n'a pas changé !

en gros, dans le premier cas, on a décidé par hypothèse que $p > 0{,}4$ était impossible. On a alors plus d'information, d'où un test qui "rejette plus facilement" $H_0$.

Il faut retenir que

la conclusion d'un test dépend des deux hypothèses

# Rédaction en sept points : alcool neutre ou nocif

(voir page 69)

1. **le modèle** : les 30 individus observés effectuent de manière indépendante une expérience de Bernoulli de paramètre $p$ (représentant la probabilité qu'un individu en état d'ébriété a de réussir l'épreuve). On note $R$ le nombre de succès.
2. **les hypothèses** : $H_0$ : $p = 0{,}4$ et $H_1$ : $p < 0{,}4$.
3. **la variable de test** : on utilise la variable $R$.
4. **la forme de la zone de rejet** : comme $R$ est plutôt plus petit sous l'hypothèse $H_0$ que sous l'hypothèse $H_1$, on va rejeter $H_0$ si $N$ est "trop petit". La zone de rejet est donc de la forme $[R \leq r_0]$.
5. **le seuil de la zone de rejet** : au niveau $\alpha = 5\%$, on trouve (voir page 69) que le plus petit $r_0$ tel que $\mathbb{P}_{H_0}(R \leq r_0) \leq \alpha$ est $r_0 = 7$.
6. **la conclusion** : puisqu'on a observé $R = 7$, on rejette $H_0$.
7. **la $p$-valeur** : $\mathbb{P}_{H_0}(R \leq 7) \simeq 4{,}3\%$ donc la $p$-valeur vaut 4,3%.

# Tests sur le paramètre d'une binomiale : résumé

Si l'on veut tester au niveau $\alpha$ le paramètre $p$ d'une binomiale $N \sim \mathcal{B}(n; p)$ pour laquelle $n$ est connu, dans tous les cas on utilise la variable de test $N$.

il y a trois possibilités :

- pour un test $H_0 : p = p_0$ contre $H_1 : p < p_0$ (où $p_0$ est connu)
  - $N$ sera plutôt plus petit sous $H_1$ que sous $H_0$, donc
  - on rejette $H_0$ si $N$ est "trop petit" : la zone de rejet est de la forme $[N \leq n]$
  - le $n$ cherché est le plus grand entier tel que $\mathbb{P}_{H_0}(N \leq n) \leq \alpha$
  - la $p$-valeur est $\mathbb{P}_{H_0}(N \leq N_{\text{obs}})$
- pour un test $H_0 : p = p_0$ contre $H_1 : p > p_0$ (où $p_0$ est connu)
  - $N$ sera plutôt plus grand sous $H_1$ que sous $H_0$, donc
  - on rejette $H_0$ si $N$ est "trop grand" : la zone de rejet est de la forme $[N \geq n]$
  - le $n$ cherché est le plus petit entier tel que $\mathbb{P}_{H_0}(N \geq n) \leq \alpha$
  - la $p$-valeur est $\mathbb{P}_{H_0}(N \geq N_{\text{obs}})$

# Tests sur le paramètre d'une binomiale : résumé

- pour un test $H_0 : p = p_0$ contre $H_1 : p \neq p_0$ (où $p_0$ est connu)
  - $N$ sera plutôt soit plus petit, soit plus grand sous $H_1$ que sous $H_0$, donc
  - on rejette $H_0$ si $N$ est "trop petit" ou "trop grand" : la zone de rejet est de la forme $[N \leq n_1 \text{ ou } N \geq n_2]$
  - $n_1$ est le plus grand entier tel que $\mathbb{P}_{H_0}(N \leq n_1) \leq \alpha/2$
  - $n_2$ est le plus petit entier tel que $\mathbb{P}_{H_0}(N \geq n_2) \leq \alpha/2$
  - la $p$-valeur est $2 \times \min\left(\mathbb{P}_{H_0}(N \leq N_{\text{obs}}), \mathbb{P}_{H_0}(N \geq N_{\text{obs}})\right)$

**Remarques :**

- vous devrez rédiger un exercice de ce type en respectant la structure en sept points
- un test sur une variable qui suit une autre loi discrète (par exemple une loi de Poisson) se fait suivant les mêmes principes

# Chapitre 2 : exercices d'autoévaluation

1. Rédiger l'exemple du test sur l'alcool (dans les deux cas pour lesquels ça n'a pas été fait ci-dessus) en sept points.
2. Traiter l'exemple des alligators au niveau $\alpha = 2\%$, en le rédigeant en sept points.

Les réponses se trouvent dans les pages suivantes.

# Rédaction en sept points : alcool neutre ou bénéfique

(voir page 70)

1. **le modèle** : les 30 individus observés effectuent de manière indépendante une expérience de Bernoulli de paramètre $p$ (représentant la probabilité qu'un individu en état d'ébriété a de réussir l'épreuve). On note $R$ le nombre de succès.
2. **les hypothèses** : $H_0$ : $p = 0{,}4$ et $H_1$ : $p > 0{,}4$.
3. **la variable de test** : on utilise la variable $R$.
4. **la forme de la zone de rejet** : comme $R$ est plutôt plus grand sous l'hypothèse $H_1$ que sous l'hypothèse $H_0$, on va rejeter $H_0$ si $R$ est "trop grand". La zone de rejet est donc de la forme $[R \geq r_0]$.
5. **le seuil de la zone de rejet** : au niveau $\alpha = 5\%$, on trouve (voir les calculs en page 70) que le plus petit $r_0$ tel que $\mathbb{P}_{H_0}(R \geq r_0) \leq \alpha$ est $r_0 = 17$.
6. **la conclusion** : puisqu'on a observé $R = 7$, on ne rejette pas $H_0$ au niveau 5%.
7. **la $p$-valeur** : $\mathbb{P}_{H_0}(R \geq 7) \simeq 98{,}3\%$ (voir les calculs en page 70) donc la $p$-valeur vaut 98,3%.

# Rédaction en sept points : alcool neutre ou actif

(voir pages 72 et 73)

1. **le modèle** : les 30 individus observés effectuent de manière indépendante une expérience de Bernoulli de paramètre $p$ (représentant la probabilité qu'un individu en état d'ébriété a de réussir l'épreuve). On note $R$ le nombre de succès.
2. **les hypothèses** : $H_0$ : $p = 0{,}4$ et $H_1$ : $p \neq 0{,}4$.
3. **la variable de test** : on utilise la variable $R$.
4. **la forme de la zone de rejet** : comme $R$ est plutôt plus grand ou bien plutôt plus petit sous l'hypothèse $H_1$ que sous l'hypothèse $H_0$, on va rejeter $H_0$ si $N$ est "trop grand ou trop petit". La zone de rejet est donc de la forme $[R \leq r_1 \text{ ou } R \geq r_2]$.
5. **le seuil de la zone de rejet** : au niveau $\alpha = 5\%$, on trouve (voir les calculs en page 72) $r_1 = 6$ et $r_2 = 18$.
6. **la conclusion** : puisqu'on a observé $R = 7$, on ne rejette pas $H_0$ au niveau 5%.
7. **la $p$-valeur** : $\mathbb{P}_{H_0}(R \leq 7) \simeq 4{,}3\%$ et $\mathbb{P}_{H_0}(R \geq 7) = 1 - \mathbb{P}_{H_0}(R \leq 6)$ vaut environ 98,3% (voir les calculs en page 73) donc la $p$-valeur vaut $2 \times 4{,}3\% = 8{,}6\%$.

# Rédaction en sept points : les alligators

1. **le modèle** : les 20 œufs observés donnent de manière indépendante un mâle avec probabilité $p$ et une femelle avec probabilité $1-p$. On note $N$ le nombre de mâles qui naissent des 20 œufs ; cette variable suit une loi $\mathcal{B}(20; p)$.
2. **les hypothèses** : $H_0$ : $p = 0{,}5$ et $H_1$ : $p > 0{,}5$.
3. **la variable de test** : on utilise la variable $N$.
4. **la forme de la zone de rejet** : comme $N$ est plutôt plus grand sous l'hypothèse $H_1$ que sous l'hypothèse $H_0$, on va rejeter $H_0$ si $N$ est "trop grand". La zone de rejet est donc de la forme $[N \geq n_0]$.
5. **le seuil de la zone de rejet** : au niveau $\alpha = 2\%$, on cherche le plus petit $n_0$ tel que $\mathbb{P}(N \geq n_0) \leq 2\%$ si $N \sim \mathcal{B}(20; 0{,}5)$. Comme $\mathbb{P}(N \geq n_0) = 1 - \mathbb{P}(N < n_0) = 1 - \mathbb{P}(N \leq n_0 - 1)$, on a $\mathbb{P}(N \geq n_0) \leq 2\% \Leftrightarrow \mathbb{P}(N \leq n_0 - 1) \geq 98\%$. D'après les tables, le plus petit $n_0$ tel que $\mathbb{P}(N \leq n_0 - 1) \geq 98\%$ vérifie $n_0 - 1 = 14$ donc $n_0 = 15$.
6. **la conclusion** : puisqu'on a observé $N = 14$, on ne rejette pas $H_0$ au niveau 2%.
7. **la $p$-valeur** : $\mathbb{P}_{H_0}(N \geq 14) = 1 - \mathbb{P}(N \leq 13) \simeq 5{,}8\%$ (voir les calculs en page 70) donc la $p$-valeur vaut 5,8%.

# Tests sur la moyenne d'une population normale (de variance connue)

# Pour quoi faire ?

Les résultats de cette section permettent de traiter l'exemple numéro 2 :

Vous arrivez en Norvège, et les dix premiers hommes (jeunes adultes) que vous rencontrez mesurent 180, 185, 184, 188, 182, 187, 182, 185, 181, 186 cm. Sachant que la taille moyenne des Français de la même catégorie est de 179 cm, doit-on en conclure que les Norvégiens sont en moyenne plus grands que les Français ?

Ou presque ! on va devoir supposer que l'on connaît la valeur d'un certain paramètre. On apprendra à s'en passer au chapitre suivant.

# Modélisation par une loi normale

Quand on s'intéresse à

- une caractéristique à valeurs continues,
- qui dépend d'un ensemble de facteurs indépendants,
- pour un individu choisi au hasard dans dans une population homogène

on modélise souvent cette caractéristique par une variable aléatoire de loi normale.

# Modélisation par une loi normale

Par exemple, on modélise la taille d'un Français (homme jeune adulte) choisi au hasard par une variable aléatoire $X$ de loi normale

car la taille est une caractéristique

- à valeurs continues,
- qui dépend de la génétique, de l'alimentation, de l'activité. . .
- et la population considérée est assez homogène.

Qu'est-ce qu'une loi normale ?

# Rappels sur les lois normales

Une variable aléatoire de loi normale est une variable :

- à densité,
- qui dépend de deux paramètres $m$ et $\sigma$.

Attention au $\sigma$ vs. $\sigma^2$ ! Si nous écrivons $X \sim \mathcal{N}(179; 49)$, nous voulons dire que c'est $\sigma^2$ qui vaut 49.

On note

$$X \sim \mathcal{N}(m; \sigma^2)$$

le fait qu'une variable aléatoire $X$ suive la loi normale de paramètres $m$ et $\sigma$.

# Rappels : variable à densité

Une variable aléatoire $X$ est "à densité" s'il existe une fonction $f$ telle que

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f(x)\,\mathrm{d}x$$

pour tous $a < b$.

Cette fonction $f$ est appelée... la densité de $X$.

# Rappels : variable à densité

Si $X$ est à densité, alors par exemple

$$\mathbb{P}(a \leq X \leq b) = \mathbb{P}(a < X \leq b) = \mathbb{P}(a \leq X < b) = \mathbb{P}(a < X < b)$$

(plus généralement on n'a jamais à faire attention à $<$ vs. $\leq$ ou $>$ vs. $\geq$)

Exemples de variables à densité :

- les variables de loi normale,
- les variables de loi de Student,
- les variables de loi $\chi^2$

(autrement dit toutes celles dont on va parler à partir de maintenant)

# Rappels sur les lois normales

La densité d'une loi normale $\mathcal{N}(m; \sigma^2)$ de paramètres $m$ et $\sigma^2$ est :

$$f_{m;\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{(x-m)^2}{2\sigma^2}$$

donc on a pour tous $a < b$

$$\mathbb{P}(a \leq X \leq b) = \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{(x-m)^2}{2\sigma^2} \, \mathrm{d}x$$

mais *on ne peut pas* calculer cette intégrale

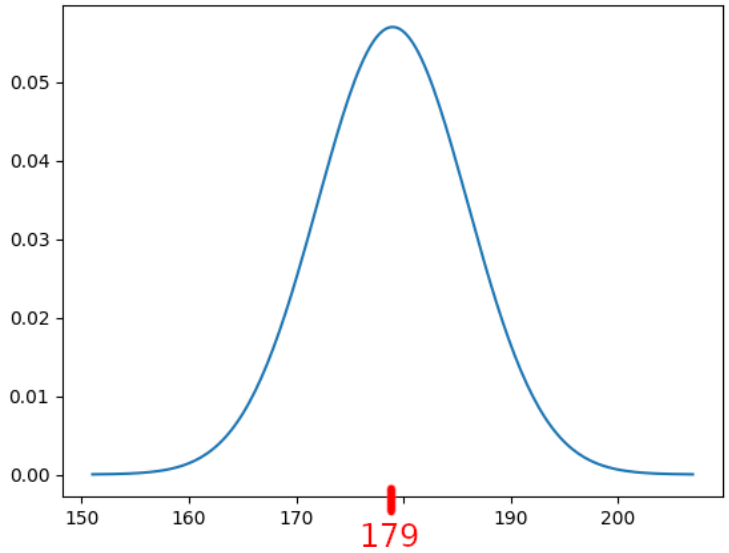
(sauf pour certains $a$ et $b$)

(sauf de manière approchée, par ordinateur)

# Rappels sur les lois normales, dépendance en $m$ et $\sigma$

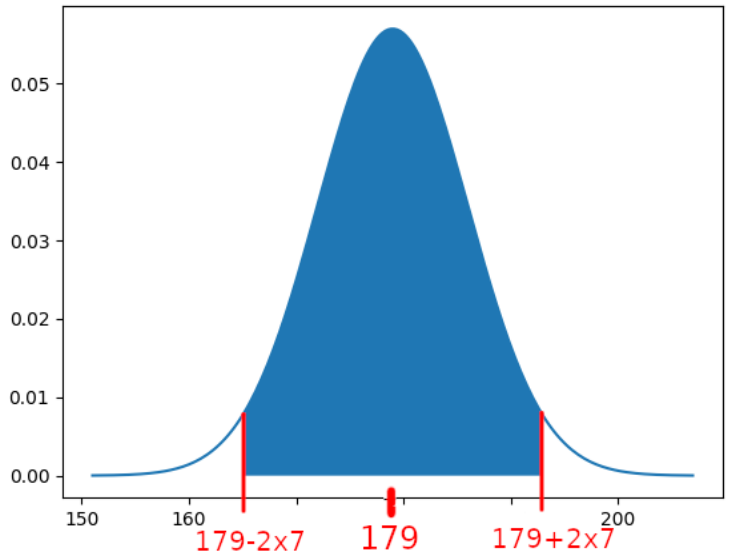
(on prendra l'exemple $m = 179$ et $\sigma = 7$)

une variable $X \sim \mathcal{N}(m; \sigma^2)$ prend des valeurs centrées autour de $m$




et prend "la majorité" ($\simeq$ 95,5%) de ses valeurs sur $[m - 2\sigma, m + 2\sigma]$ :

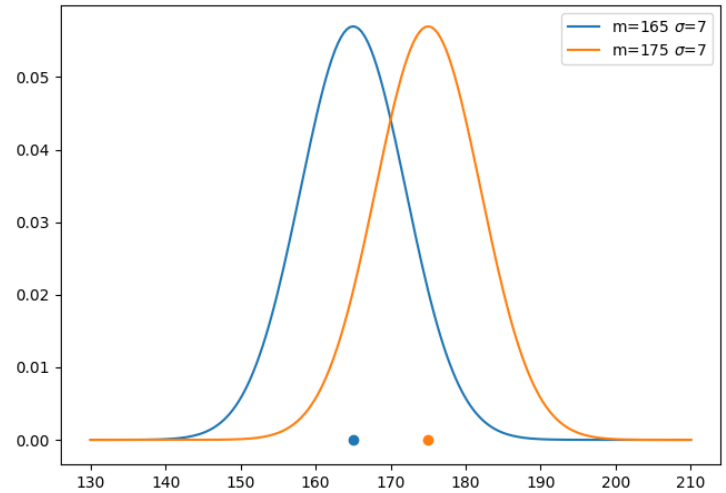
$$\int_{m-2\sigma}^{m+2\sigma} f_{m,\sigma^2}(x)\,\mathrm{d}x \simeq 0{,}955$$



# Rappels sur les lois normales, dépendance en $m$ et $\sigma$

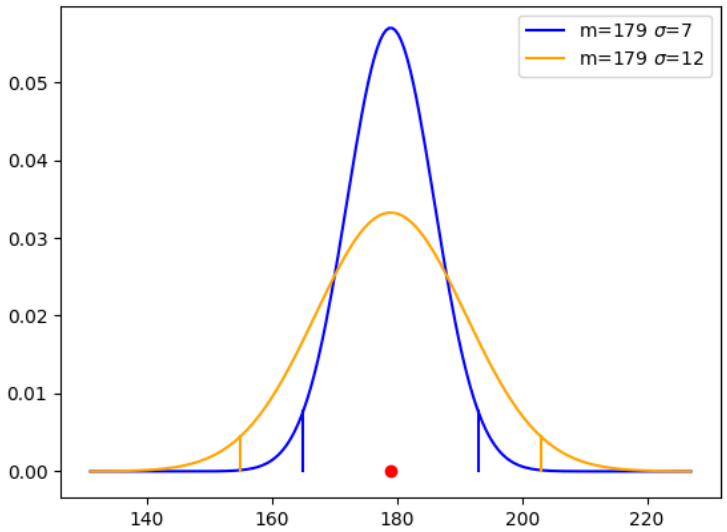
à $\sigma$ fixé, les valeurs probables sont plus élevées quand $m$ augmente

(les valeurs de $m$ sont indiquées)




à $m$ fixé, la variable prend des valeurs plus dispersées quand $\sigma$ augmente

(les valeurs de $m$ et $m \pm 2\sigma$ sont indiquées)



# Rappels : loi normale, espérance et variance

pour une normale $X \sim \mathcal{N}(m; \sigma^2)$ :

l'espérance, qui représente la "moyenne pondérée" des valeurs possibles, est

$$\mathbb{E}(X) \overset{\text{déf}}{=} \int x \, f_{m,\sigma^2}(x) \, \mathrm{d}x \quad \text{vaut} \quad \mathbb{E}(X) = m$$

la variance, qui donne une information sur la "dispersion" des valeurs possibles, est

$$\mathbb{V}(N) \overset{\text{déf}}{=} \int \big(x - \mathbb{E}(X)\big)^2 \, f_{m,\sigma^2}(x) \, \mathrm{d}x \quad \text{vaut} \quad \mathbb{V}(X) = \sigma^2$$

(ne pas confondre l'écart-type $\sigma$ et la variance $\sigma^2$)

# Modélisation par une loi normale

On modélise la taille $X$ d'un Français (homme jeune adulte) choisi au hasard, exprimée en centimètres, par une variable aléatoire de loi normale $\mathcal{N}(m; \sigma^2)$ avec :

$$m = 179, \qquad \sigma = 7$$

mais alors $\mathbb{P}(X < 0) \neq 0$, donc le modèle donne une probabilité non nulle de choisir une taille négative ?

oui, mais

- la proba en question $\simeq 1{,}6 \times 10^{-144}$ est tellement petite que ce n'est pas grave,
- les normales sont tellement pratiques par ailleurs qu'on les utilise quand même.

# Pourquoi tant de normales ?

En particulier à cause du *théorème central limite* :

si $X_1, \ldots, X_n$ sont des variables aléatoires indépendantes et qui suivent la même distribution statistique, alors "si $n$ est grand"

$$\frac{X_1 + \ldots + X_n - n\,\mathbb{E}(X_1)}{\sqrt{n}} \text{ suit à peu près une loi } \mathcal{N}\big(0, \mathbb{V}(X_1)\big)$$

"donc" toute variable qui résulte de l'accumulation d'un grand nombre de variables *indépendantes* et *d'amplitudes comparables* suit à peu près une loi normale.

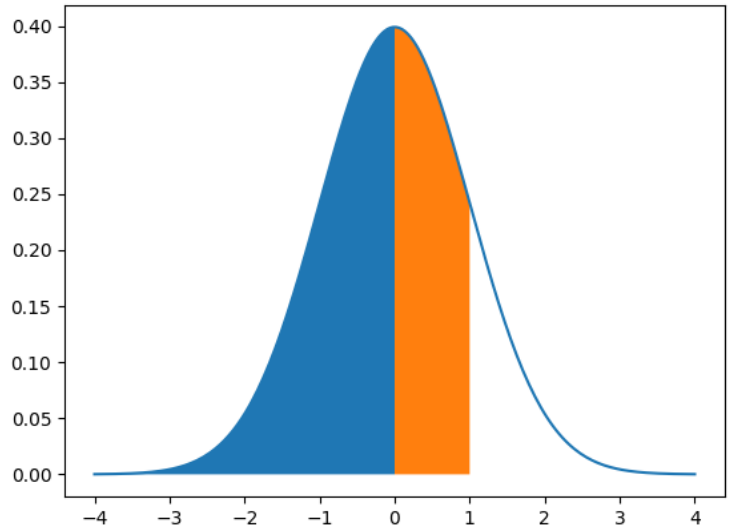
# Rappels : tables numériques pour les normales

Vous avez une seule table concernant les lois normales.

Cette table donne les valeurs de la *fonction de répartition* $t \mapsto \mathbb{P}(U \leq t)$

- pour $U$ de loi $\mathcal{N}(0;1)$
- pour $t \geq 0$.

Remarquez que si $t \geq 0$ alors
$\mathbb{P}(U \leq t) \geq \mathbb{P}(U \leq 0) = 0{,}5$




(on essaiera de se tenir à la convention qu'une variable $U$ suit une loi $\mathcal{N}(0;1)$)

# Rappels : tables numériques pour les normales

Par conséquent, il est facile de :

- calculer une probabilité $\mathbb{P}(U \leq t)$ pour $U \sim \mathcal{N}(0;1)$ et $t \geq 0$ : on lit directement la valeur. Par exemple, $\mathbb{P}(X \leq 0{,}53) = 0{,}7019$ :

| $t$ | 0 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 |
|---|---|---|---|---|---|---|
| 0 | 0.5000 | 0.5040 | 0.5080 | 0.5120 | 0.5160 | 0.5199 |
| 0.1 | 0.5398 | 0.5438 | 0.5478 | 0.5517 | 0.5557 | 0.5596 |
| 0.2 | 0.5793 | 0.5832 | 0.5871 | 0.5910 | 0.5948 | 0.5987 |
| 0.3 | 0.6179 | 0.6217 | 0.6255 | 0.6293 | 0.6331 | 0.6368 |
| 0.4 | 0.6554 | 0.6591 | 0.6628 | 0.6664 | 0.6700 | 0.6736 |
| 0.5 | 0.6915 | 0.6950 | 0.6985 | 0.7019 | 0.7054 | 0.7088 |
| 0.6 | 0.7257 | 0.7291 | 0.7324 | 0.7357 | 0.7389 | 0.7422 |

- trouver le plus grand $t$ tel que $\mathbb{P}(U \leq t) \leq 1-\alpha$ si $1-\alpha \geq 0{,}5$ : on peut lire la valeur en parcourant la table, par exemple le plus grand $t$ tel $\mathbb{P}(U \leq t) \leq 0{,}73$ est $t_0 = 0{,}61$ :

| $t$ | 0 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 |
|---|---|---|---|---|---|---|
| 0 | 0.5000 | 0.5040 | 0.5080 | 0.5120 | 0.5160 | 0.5199 |
| 0.1 | 0.5398 | 0.5438 | 0.5478 | 0.5517 | 0.5557 | 0.5596 |
| 0.2 | 0.5793 | 0.5832 | 0.5871 | 0.5910 | 0.5948 | 0.5987 |
| 0.3 | 0.6179 | 0.6217 | 0.6255 | 0.6293 | 0.6331 | 0.6368 |
| 0.4 | 0.6554 | 0.6591 | 0.6628 | 0.6664 | 0.6700 | 0.6736 |
| 0.5 | 0.6915 | 0.6950 | 0.6985 | 0.7019 | 0.7054 | 0.7088 |
| 0.6 | 0.7257 | 0.7291 | 0.7324 | 0.7357 | 0.7389 | 0.7422 |

# Rappels : tables numériques pour les normales

Comment faire si l'on veut calculer une probabilité

- concernant $X \sim \mathcal{N}(m; \sigma^2)$ avec $m \neq 0$ et/ou $\sigma \neq 1$ ?
- du type $\mathbb{P}(X \leq t)$ pour $t < 0$ ?
- d'un autre type que $\mathbb{P}(X \leq t)$, par exemple $\mathbb{P}(X \geq t)$ ?

# Rappels sur les lois normales : centrer et réduire

On a

$$X \sim \mathcal{N}(m;\sigma^2) \quad \Longleftrightarrow \quad \frac{X-m}{\sigma} \sim \mathcal{N}(0,1)$$

Donc si $X \sim \mathcal{N}(m;\sigma^2)$, on a par exemple :

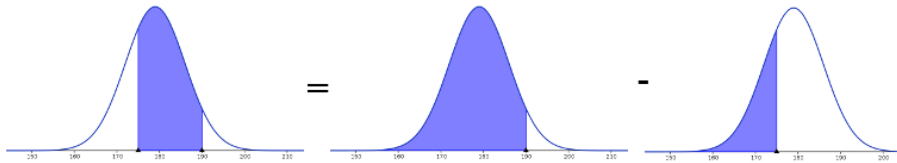
$$\mathbb{P}(a \leq X \leq b) = \mathbb{P}(\frac{a-m}{\sigma} \leq \frac{X-m}{\sigma} \leq \frac{b-m}{\sigma}) = \mathbb{P}(\frac{a-m}{\sigma} \leq U \leq \frac{b-m}{\sigma})$$

où $U \sim \mathcal{N}(0,1)$

(la loi $\mathcal{N}(0;1)$ est appelée *normale centrée réduite*)

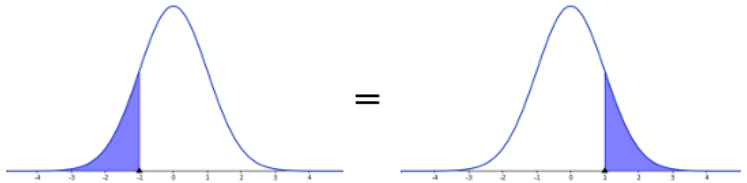
# Rappels sur les lois normales : règles de calcul

$$\mathbb{P}(a \leq U \leq b) = \mathbb{P}(U \leq b) - \mathbb{P}(U \leq a)$$



$$\mathbb{P}(U \geq b) = 1 - \mathbb{P}(U \leq b)$$

$$\mathbb{P}(U \leq -t) = \mathbb{P}(U \geq +t)$$



(les deux premières sont vraies aussi pour $X \sim \mathcal{N}(m; \sigma^2)$, pas la troisième)

# Rappels sur les lois normales : exemple de calcul

On veut calculer $\mathbb{P}(8 \leq X \leq 13)$ pour $X \sim \mathcal{N}(10,4)$

- on centre comme expliqué en page 98 :

$$\mathbb{P}(8 \leq X \leq 13) = \mathbb{P}(\frac{8-10}{2} \leq \frac{X-10}{2} \leq \frac{13-10}{2})$$
$$= \mathbb{P}(-1 \leq U \leq \frac{3}{2}) \text{ pour } U \sim \mathcal{N}(0,1)$$

- on applique la première règle page 99 :

$$\mathbb{P}(-1 \leq U \leq \frac{3}{2}) = \mathbb{P}(U \leq \frac{3}{2}) - \mathbb{P}(U \leq -1)$$

- on calcule $\mathbb{P}(U \leq \frac{3}{2}) \simeq 0{,}9332$ par lecture directe comme en page 96
- on calcule $\mathbb{P}(U \leq -1)$ en appliquant les troisième puis deuxième règles page 99 :

$$\mathbb{P}(U \leq -1) = \mathbb{P}(U \geq 1) = 1 - \mathbb{P}(U \leq 1)$$

et on trouve $\mathbb{P}(U \leq 1) = 0{,}8413$ par lecture directe comme en page 96

- pour finir :

$$\mathbb{P}(8 \leq X \leq 13) \simeq 0{,}9332 - (1 - 0{,}8413) \simeq 0{,}7745$$

# Test sur la moyenne d'*une* normale (variance connue)

Vous arrivez en Norvège, et le premier homme jeune adulte que vous rencontrez mesure 188 cm. Sachant que la taille moyenne des Français de la même catégorie est de 179 cm, avec un écart type de 7 cm, doit-on en conclure que les Norvégiens sont en moyenne plus grands que les Français ?

(on admettra que l'écart-type est le même chez les Norvégiens et chez les Français, et on fera le test au niveau $\alpha = 3\%$)

# Test sur la moyenne d'*une* normale (variance connue)

On se base sur l'analyse donnée en page 24, mais on écrit directement le tout en sept points.

1. **le modèle** : la taille $X$ du Norvégien observé suit une loi $\mathcal{N}(m; 7^2)$.
2. **les hypothèses** : $\mathrm{H}_0$ : $m = 179$ et $\mathrm{H}_1$ : $m > 179$.
3. **la variable de test** : on utilise la variable $X$.
4. **la forme de la zone de rejet** : comme $X$ est plutôt plus grand sous l'hypothèse $\mathrm{H}_1$ que sous l'hypothèse $\mathrm{H}_0$, on va rejeter $\mathrm{H}_0$ si $X$ est "trop grand". La zone de rejet est donc de la forme $[X \geq x_0]$.
5. **le seuil de la zone de rejet** : au niveau $\alpha = 3\%$, on cherche le plus petit $x_0$ tel que $\mathbb{P}(X \geq x_0) \leq 3\%$ si $X \sim \mathcal{N}(179; 7^2)$. Il y a un peu de calculs.

# Test sur la moyenne d'*une* normale (variance connue)

On cherche le plus petit $x_0$ tel que $\mathbb{P}(X \geq x_0) \leq 3\%$ si $X \sim \mathcal{N}(179; 7^2)$

- on centre comme expliqué en page 98 :

$$\mathbb{P}(X \geq x_0) = \mathbb{P}(\frac{X-179}{7} \geq \frac{x_0-179}{7}) = \mathbb{P}(U \geq \frac{x_0-179}{7}) \text{ pour } U \sim \mathcal{N}(0,1)$$

- on applique la deuxième règle page 99 :

$$\mathbb{P}(U \geq \frac{x_0-179}{7}) = 1 - \mathbb{P}(U \leq \frac{x_0-179}{7})$$

on cherche donc le plus petit $x_0$ tel que

$$\mathbb{P}(U \leq \frac{x_0-179}{7}) \geq 1 - 0{,}03$$

et comme $1 - 0{,}03$ est supérieur à 0,5, on peut directement lire cette quantité dans la table de valeurs numériques : on a $\frac{x_0-179}{7} = 1{,}89$. On trouve donc pour finir que le seuil vaut $x_0 = 179 + 1{,}89 \times 7 \simeq 192{,}2$z

# Test sur la moyenne d'*une* normale (variance connue)

6. **la conclusion** : puisqu'on a observé $X = 188$ qui est inférieur à 192,2 on ne rejette pas $H_0$ au niveau 3%.
7. **la *p*-valeur** : la zone de rejet étant de la forme $[X \geq x_0]$ et l'observation étant $X = 188$ la *p*-valeur vaut $\mathbb{P}_{H_0}(X \geq 188)$. Encore un coup il y a un peu de calculs :
   - on centre comme expliqué en page 98 :
   $$\mathbb{P}(X \geq 188) = \mathbb{P}(\frac{X - 179}{7} \geq \frac{188 - 179}{7}) = \mathbb{P}(U \geq \frac{9}{7}) \text{ pour } U \sim \mathcal{N}(0,1)$$
   - on applique la deuxième règle page 99 :
   $$\mathbb{P}(U \geq \frac{9}{7}) = 1 - \mathbb{P}(U \leq \frac{9}{7})$$

   et $9/7 \simeq 1{,}29$. On lit directement dans la table $\mathbb{P}(U \leq 1{,}29) \simeq 0{,}9015$ d'où une *p*-valeur de $1 - 0{,}9015 = 9{,}85\%$.

# Test sur la moyenne d'une *population* normale

Vous arrivez en Norvège, et les dix premiers hommes (jeunes adultes) que vous rencontrez mesurent 180, 185, 184, 188, 182, 187, 182, 185, 181, 186 cm. Sachant que la taille moyenne des Français de la même catégorie est de 179 cm, avec un écart type de 7 cm, doit-on en conclure que les Norvégiens sont en moyenne plus grands que les Français ?

(on admettra que l'écart-type est le même chez les Norvégiens et chez les Français, et on fera le test au niveau $\alpha = 3\%$)

peut-on obtenir un test plus efficace avec ces dix observations supposées indépendantes qu'avec une seule observation ?

# Rappels

Si $X \sim \mathcal{N}(m,\sigma^2)$ et $\lambda \in \mathbb{R}$ alors

$$\lambda X \sim \mathcal{N}(\lambda m,\lambda^2\sigma^2)$$

Si $X_1 \sim \mathcal{N}(m_1,\sigma_1^2)$ et $X_2 \sim \mathcal{N}(m_2,\sigma_2^2)$ sont indépendants alors

$$X_1 + X_2 \sim \mathcal{N}(m_1 + m_2,\sigma_1^2 + \sigma_2^2)$$

Une conséquence est que

si $X_1,\ldots,X_n$ sont indépendants et de même loi $\mathcal{N}(m;\sigma^2)$ alors la *moyenne empirique*

$$\overline{X}_n \stackrel{\text{déf}}{=} \frac{X_1 + \ldots + X_n}{n} \text{ suit une loi } \mathcal{N}(m,\frac{\sigma^2}{n}).$$

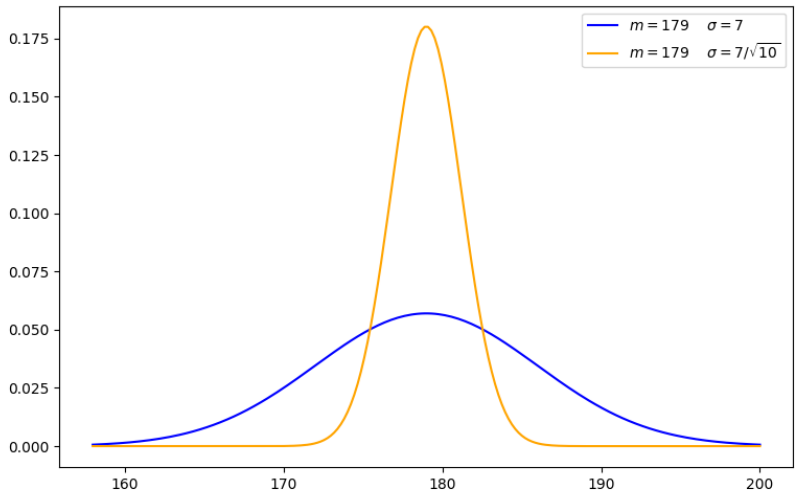
(une famille $X_1,\ldots,X_n$ de variables aléatoires indépendantes et qui suivent la même distribution statistique est appelée un *n-échantillon*).

# Test sur la moyenne d'une *population* normale

Par conséquent, si l'on note $X_1, \ldots, X_{10}$ les tailles de dix Norvégiens choisis au hasard, alors la moyenne empirique

$$\overline{X}_{10} = \frac{X_1 + \ldots + X_{10}}{10} \text{ suit une loi } \mathcal{N}(179, \frac{7^2}{10}).$$

Donc $\overline{X}_{10}$ (moyenne des tailles de dix Norvégiens observés) a la même moyenne que $X_1$ (taille d'un seul Norvégien observé) mais une variance dix fois inférieure. La dispersion de l'observation est donc bien moindre, on devrait pouvoir faire des tests plus efficaces.



# Test sur la moyenne d'une *population* normale (variance connue)

On reprend le test avec les dix observations.

1. **le modèle** : les tailles $X_1, \ldots, X_{10}$ des dix Norvégiens observés forment un 10-échantillon de loi $\mathcal{N}(m; 7^2)$.
2. **les hypothèses** : $H_0$ : $m = 179$ et $H_1$ : $m > 179$.
3. **la variable de test** : on utilise la moyenne empirique $\overline{X}_{10} = \frac{X_1 + \ldots + X_{10}}{10}$.
4. **la forme de la zone de rejet** : comme $\overline{X}_{10}$ est plutôt plus grand sous l'hypothèse $H_1$ que sous l'hypothèse $H_0$, on va rejeter $H_0$ si $\overline{X}_{10}$ est "trop grand". La zone de rejet est donc de la forme $[\overline{X}_{10} \geq x_0']$.
5. **le seuil de la zone de rejet** : au niveau $\alpha = 3\%$, on cherche le plus petit $x_1$ tel que $\mathbb{P}(\overline{X}_{10} \geq x_1) \leq 3\%$ si $\overline{X}_{10} \sim \mathcal{N}(179; 7^2)$. Il y a un peu de calculs.

# Test sur la moyenne d'une *population* normale (variance connue)

On cherche le plus petit $x_1$ tel que $\mathbb{P}(\overline{X}_{10} \geq x_1) \leq 3\%$ si $\overline{X}_{10} \sim \mathcal{N}(179; \frac{7^2}{10})$

- on centre comme expliqué en page 98 :

$$\mathbb{P}(X \geq x_1) = \mathbb{P}(\frac{X-179}{7/\sqrt{10}} \geq \frac{x_1-179}{7/\sqrt{10}}) = \mathbb{P}(U \geq \frac{x_1-179}{7/\sqrt{10}}) \text{ pour } U \sim \mathcal{N}(0,1)$$

- on applique la deuxième règle page 99 :

$$\mathbb{P}(U \geq \frac{x_1-179}{7/\sqrt{10}}) = 1 - \mathbb{P}(U \leq \frac{x_1-179}{7/\sqrt{10}})$$

on cherche donc le plus petit $x_1$ tel que

$$\mathbb{P}(U \leq \frac{x_1-179}{7/\sqrt{10}}) \geq 1 - 0{,}03$$

et comme $1-0{,}03$ est supérieur à 0,5, on peut directement lire cette quantité dans la table de valeurs numériques : on a $\frac{x_1-179}{7/\sqrt{10}} = 1{,}89$. On trouve donc pour finir que le seuil vaut $x_1 = 179 + 1{,}89 \times \frac{7}{\sqrt{10}} \simeq 183{,}2$.

# Test sur la moyenne d'une *population* normale (variance connue)

6. **la conclusion** : on observe $\overline{X}_{10} = \frac{180+185+184+188+182+187+182+185+181+186}{10} = 184$ qui est supérieur à 183,2. On rejette $H_0$ au niveau 3%.
7. **la *p*-valeur** : la zone de rejet étant de la forme $[X \geq x_0]$ et l'observation étant $X = 188$ la *p*-valeur vaut $\mathbb{P}_{H_0}(X \geq 184)$. Encore un coup il y a un peu de calculs :
   - on centre comme expliqué en page 98 :
   $$\mathbb{P}(X \geq 184) = \mathbb{P}\left(\frac{X-179}{7/\sqrt{10}} \geq \frac{184-179}{7/\sqrt{10}}\right) = \mathbb{P}\left(U \geq \frac{5\sqrt{10}}{7}\right) \text{ pour } U \sim \mathcal{N}(0,1)$$
   - on applique la deuxième règle page 99 :
   $$\mathbb{P}\left(U \geq \frac{5\sqrt{10}}{7}\right) = 1 - \mathbb{P}\left(U \leq \frac{5\sqrt{10}}{7}\right)$$

   et $5\sqrt{10}/7 \simeq 2{,}26$. On lit $\mathbb{P}(U \leq 2{,}26) \simeq 0{,}9881$ dans la table, d'où une *p*-valeur de $1 - 0{,}9881 = 1{,}19\%$.

# Test sur la moyenne d'une *population* normale (variance connue)

On va traiter le cas des Boliviens et des Italiens :

- Vous arrivez en Bolivie, et les dix premiers hommes (jeunes adultes) que vous rencontrez mesurent 160, 165, 164, 168, 162, 167, 162, 165, 161, 166 cm. Sachant que la taille moyenne des Français de la même catégorie est de 179 cm, avec un écart type de 7 cm, doit-on en conclure que les Boliviens sont en moyenne plus grands que les Français ?
- Vous arrivez en Italie, et les dix premiers hommes (jeunes adultes) que vous rencontrez mesurent 173, 180, 176, 190, 185, 188, 164, 183, 173, 175 cm. Sachant que la taille moyenne des Français de la même catégorie est de 179 cm, avec un écart type de 7 cm, doit-on en conclure que les Boliviens sont en moyenne plus grands que les Français ?

(on admettra que l'écart-type est le même chez les Boliviens ou Italiens et chez les Français, et on fera le test au niveau $\alpha = 3\%$)

# Test sur la moyenne d'une *population* normale (variance connue)

On reprend le test dans le cas des Boliviens.

1. **le modèle** : les tailles $X_1, \ldots, X_{10}$ des dix Boliviens observés forment un 10-échantillon de loi $\mathcal{N}(m; 7^2)$.
2. **les hypothèses** : $H_0$ : $m = 179$ et $H_1{}'$ : $m < 179$.
3. **la variable de test** : on utilise la variable $\overline{X}_{10} = \frac{X_1 + \ldots + X_{10}}{10}$.
4. **la forme de la zone de rejet** : comme $\overline{X}_{10}$ est plutôt plus petit sous l'hypothèse $H_1'$ que sous l'hypothèse $H_0$, on va rejeter $H_0$ si $\overline{X}_{10}$ est "trop petit". La zone de rejet est donc de la forme $[\overline{X}_{10} \leq x_2]$.
5. **le seuil de la zone de rejet** : au niveau $\alpha = 3\%$, on cherche le plus grand $x_2$ tel que $\mathbb{P}(\overline{X}_{10} \leq x_2) \leq 3\%$ si $\overline{X}_{10} \sim \mathcal{N}(179; 7^2)$. Il y a un peu de calculs.

# Test sur la moyenne d'une *population* normale (variance connue)

On cherche le plus grand $x_2$ tel que $\mathbb{P}(\overline{X}_{10} \leq x_2) \leq 3\%$ si $\overline{X}_{10} \sim \mathcal{N}(179; \frac{7^2}{10})$

- on centre comme expliqué en page 98 :

$$\mathbb{P}(X \leq x_2) = \mathbb{P}(\frac{X-179}{7/\sqrt{10}} \leq \frac{x_2-179}{7/\sqrt{10}}) = \mathbb{P}(U \leq \frac{x_2-179}{7/\sqrt{10}}) \text{ pour } U \sim \mathcal{N}(0,1)$$

  On cherche donc le plus grand $x_2$ tel que

$$\mathbb{P}(U \leq \frac{x_2-179}{7/\sqrt{10}}) \leq 0{,}03$$

  mais comme cette probabilité est plus petite que 0,5 on n'a aucune chance de trouver $\frac{x_2-179}{7/\sqrt{10}}$ directement dans le tableau : on a forcément $\frac{x_2-179}{7/\sqrt{10}} < 0$.

- on applique la troisième puis la deuxième règle page 99 :

$$\mathbb{P}(U \leq \frac{x_2-179}{7/\sqrt{10}}) = \mathbb{P}(U \geq -\frac{x_2-179}{7/\sqrt{10}}) = 1 - \mathbb{P}(U \leq -\frac{x_2-179}{7/\sqrt{10}})$$

  Donc si $x_2$ est "le plus grand tel que $\mathbb{P}(\overline{X}_{10} \leq x_2) \leq 3\%$" alors $-\frac{x_2-179}{7/\sqrt{10}}$ est "le plus petit tel que $\mathbb{P}(U \leq -\frac{x_2-179}{7/\sqrt{10}}) \geq 1 - 0{,}03$" (attention au signe moins). On trouve dans la table $-\frac{x_2-179}{7/\sqrt{10}} = 1{,}89$ donc $x_2 = 179 - 1{,}89 \times \frac{7}{\sqrt{10}} \simeq 174{,}82$.

# Test sur la moyenne d'une *population* normale (variance connue)

6. **la conclusion** : on observe $\overline{X}_{10} = \frac{160+165+164+168+162+167+162+165+161+166}{10} = 164$ qui est inférieur à 174,82. On rejette $H_0$ au niveau 3%.
7. **la *p*-valeur** : la zone de rejet étant de la forme $[X \leq x_2]$ et l'observation étant $X = 168$ la *p*-valeur vaut $\mathbb{P}_{H_0}(X \leq 168)$. Encore un coup il y a un peu de calculs :
   - on centre comme expliqué en page 98 :
   $$\mathbb{P}(X \leq 168) = \mathbb{P}\left(\frac{X-179}{7/\sqrt{10}} \leq \frac{168-179}{7/\sqrt{10}}\right) = \mathbb{P}\left(U \leq -\frac{11\sqrt{10}}{7}\right)$$
   - on applique la deuxième règle page 99 :
   $$\mathbb{P}\left(U \leq -\frac{11\sqrt{10}}{7}\right) = \mathbb{P}\left(U \geq \frac{11\sqrt{10}}{7}\right) = 1 - \mathbb{P}\left(U \leq \frac{11\sqrt{10}}{7}\right)$$

   et $11\sqrt{10}/7 \simeq 4{,}97$. On ne peut pas lire $\mathbb{P}(U \leq 4{,}97)$ dans la table, mais on sait que c'est supérieur à la dernière valeur disponible qui est $\mathbb{P}(U \leq 3{,}69) \simeq 0{,}9999$ d'où une *p*-valeur inférieure $1 - 0{,}9999 = 0{,}01\%$. On n'a donc aucune hésitation sur la conclusion.

# Test sur la moyenne d'une *population* normale (variance connue)

On reprend le test dans le cas des Italiens.

1. **le modèle** : les tailles $X_1, \ldots, X_{10}$ des dix Italiens observés forment un 10-échantillon de loi $\mathcal{N}(m; 7^2)$.
2. **les hypothèses** : $H_0$ : $m = 179$ et $H_1''$ : $m \neq 179$.
3. **la variable de test** : on utilise la variable $\overline{X}_{10} = \frac{X_1 + \ldots + X_{10}}{10}$.
4. **la forme de la zone de rejet** : comme $\overline{X}_{10}$ est sous l'hypothèse $H_1'$ plutôt plus petit ou plutôt plus grand que sous l'hypothèse $H_0$, on va rejeter $H_0$ si $\overline{X}_{10}$ est "trop petit ou trop grand". La zone de rejet est donc de la forme $[\overline{X}_{10} \leq x_3 \text{ ou } \overline{X}_{10} \geq x_4]$.
5. **le seuil de la zone de rejet** : au niveau $\alpha = 3\%$, on cherche le plus grand $x_3$ tel que $\mathbb{P}(\overline{X}_{10} \leq x_3) \leq 1{,}5\%$ et le plus petit $x_4$ tel que $\mathbb{P}(\overline{X}_{10} \geq x_4) \leq 1{,}5\%$ si $\overline{X}_{10} \sim \mathcal{N}(179; 7^2)$. Il y a un peu de calculs mais ils sont similaires à ceux que l'on a déjà faits. On trouve que $x_3 = 179 - 2{,}17 \times \frac{7}{\sqrt{10}} \simeq 174{,}20$ et $x_4 = 179 + 2{,}17 \times \frac{7}{\sqrt{10}} \simeq 183{,}80$.

# Test sur la moyenne d'une *population* normale (variance connue)

6. **la conclusion** : on observe $\overline{X}_{10} = \frac{173+180+176+190+185+188+164+183+173+175}{10} = 178{,}7$ qui est supérieur à $x_3$ et inférieur à $x_4$. On conserve $H_0$ au niveau 3%.
7. **la *p*-valeur** : la zone de rejet étant de la forme $[X \leq x_3 \text{ ou } X \geq x_4]$ et l'observation étant $X = 178{,}7$ la *p*-valeur vaut $2 \times \min\left(\mathbb{P}_{H_0}(X \leq 178{,}7), \mathbb{P}_{H_0}(X \geq 178{,}7)\right)$. Comme $\mathbb{P}_{H_0}(X \geq 178{,}7) = 1 - \mathbb{P}_{H_0}(X \leq 178{,}7)$ il suffit de calculer $\mathbb{P}_{H_0}(X \leq 178{,}7)$. Pour cela :
   - on centre comme expliqué en page 98 :
   $$\mathbb{P}(X \leq 178{,}7) = \mathbb{P}\left(\frac{X - 179}{7/\sqrt{10}} \leq \frac{178{,}7 - 179}{7/\sqrt{10}}\right) = \mathbb{P}\left(U \leq -\frac{0{,}3\sqrt{10}}{7}\right)$$
   - on applique la deuxième règle page 99 :
   $$\mathbb{P}\left(U \leq -\frac{0{,}3\sqrt{10}}{7}\right) = \mathbb{P}\left(U \geq \frac{0{,}3\sqrt{10}}{7}\right) = 1 - \mathbb{P}\left(U \leq \frac{0{,}3\sqrt{10}}{7}\right)$$

   et $0{,}3\sqrt{10}/7 \simeq 0{,}14$. On lit $\mathbb{P}(U \leq 0{,}14) \simeq 0{,}5557$ dans la table, donc $\mathbb{P}(X \leq 178{,}7) \simeq 0{,}4443$ et l'autre valeur est plus grande. On trouve donc une *p*-valeur de $2 \times 0{,}4443 = 88{,}86\%$.

# Une remarque utile

Vous avez peut-être observé que dans les exemples ci-dessus les seuils que l'on a trouvé sont toujours de la forme $179 \pm u \times \frac{7}{\sqrt{10}}$ où $u$ a été trouvé à partir du tableau de la normale centrée réduite.

De manière plus générale dans les tests avec $H_0 : m = m_0$ concernant un $n$-échantillon d'une loi normale $\mathcal{N}(m; \sigma^2)$ dont on connaît la variance $\sigma^2$, les seuils seront toujours de la forme

$$m_0 \pm u \times \frac{\sigma}{\sqrt{n}}.$$

On pourrait expliciter comment trouver le $u$. La seule information que je vous conseille de retenir est que dans un test bilatère, les deux seuils sont de la forme

$$m_0 - u \times \frac{\sigma}{\sqrt{n}} \text{ et } m_0 + u \times \frac{\sigma}{\sqrt{n}}.$$

Il suffit donc d'en calculer un pour avoir l'autre.

# Une autre utilisation des lois normales

Les lois normales peuvent aussi nous servir pour les calculs concernant des lois binomiales $\mathcal{B}(n; p)$ avec un $n$ grand (et $p$ pas trop petit).

(on en a déjà parlé page 62)

En pratique, si les trois conditions

$$n \geq 30 \quad np \geq 15 \quad n(1-p) \geq 15$$

sont satisfaites, alors la loi binomiale $\mathcal{B}(n; p)$ est à peu près la même chose qu'une loi normale $\mathcal{N}\big(np; np(1-p)\big)$.

(les coefficients sont tels que les deux lois ont mêmes espérance et variance)

# Une autre utilisation des lois normales

Si par exemple on avait traité le test sur l'alcool neutre ou nocif (voir page 75) mais avec $n = 100$ participants, on aurait considéré $R \sim \mathcal{B}(100; p)$ pour tester $\mathrm{H}_0 : p = 0{,}4$ contre $\mathrm{H}_1 : p < 0{,}4$.

Pour obtenir le seuil, on aurait alors cherché le plus petit $r_0'$ tel que $\mathbb{P}_{\mathrm{H}_0}(R \leq r_0') \leq \alpha$. Comme les trois conditions

$$100 \geq 15, \quad 100 \times 0{,}4 \geq 5, \quad 100 \times 0{,}6 \geq 5$$

sont vérifiées, on peut dire que sous $\mathrm{H}_0$, la variable $R$ suit à peu près une loi

$$\mathcal{N}(100 \times 0{,}4; 100 \times 0{,}4 \times 0{,}6)$$

et le plus petit $r_0'$ tel que $\mathbb{P}_{\mathrm{H}_0}(R \leq r_0') \leq \alpha$ en supposant $R \sim \mathcal{N}(40; 24)$.

Par exemple, pour $\alpha = 5\%$ on trouve $r_0' = 32$ sans utiliser l'approximation (calcul fait sur ordinateur car on n'a pas la table pour une binomiale $\mathcal{B}(100; 0{,}4)$) et $r_0' \simeq 31{,}94$ en l'utilisant (en appliquant les règles de calcul en pages 98 et 99).

# Tests sur la moyenne d'une population normale à variance connue : résumé

Si l'on veut tester au niveau $\alpha$ la moyenne $m$ d'un $n$-échantillon $X_1, \ldots, X_n$ de $\mathcal{N}(m; \sigma^2)$ où $\sigma$ est connu, dans tous les cas on utilise la variable

$$\overline{X}_n \stackrel{\text{def}}{=} \frac{X_1 + \ldots + X_n}{n}$$

il y a trois possibilités :

- pour un test $H_0 : m = m_0$ contre $H_1 : m < m_0$ (où $m_0$ est connu)
  - $\overline{X}_n$ sera plutôt plus petit sous $H_1$ que sous $H_0$, donc
  - on rejette $H_0$ si $\overline{X}_n$ est "trop petit" : la zone de rejet est de la forme $[\overline{X}_n \leq x]$
  - le $x$ cherché est le plus grand réel tel que $\mathbb{P}_{H_0}(\overline{X}_n \leq x) \leq \alpha$
  - la $p$-valeur est $\mathbb{P}_{H_0}(\overline{X}_n \leq \overline{X}_{n,\text{obs}})$
- pour un test $H_0 : m = m_0$ contre $H_1 : m > m_0$ (où $m_0$ est connu)
  - $\overline{X}_n$ sera plutôt plus grand sous $H_1$ que sous $H_0$, donc
  - on rejette $H_0$ si $\overline{X}_n$ est "trop grand" : la zone de rejet est de la forme $[\overline{X}_n \geq x]$
  - le $x$ cherché est le plus petit réel tel que $\mathbb{P}_{H_0}(\overline{X}_n \geq x) \leq \alpha$
  - la $p$-valeur est $\mathbb{P}_{H_0}(\overline{X}_n \geq \overline{X}_{n,\text{obs}})$

# Tests sur la moyenne d'une population normale à variance connue : résumé

- pour un test $H_0 : m = m_0$ contre $H_1 : m \neq m_0$ (où $m_0$ est connu)
  - $\overline{X}_n$ sera soit plus petit, soit plus grand sous $H_1$ que sous $H_0$, donc
  - on rejette $H_0$ si $\overline{X}_n$ est "trop petit ou trop grand" : la zone de rejet est de la forme $[\overline{X}_n \leq x_1$ ou $\overline{X}_n \geq x_2]$
  - $x_1$ est le plus grand réel tel que $\mathbb{P}_{H_0}(\overline{X}_n \leq x_1) \leq \alpha/2$
  - $x_2$ est le plus petit réel tel que $\mathbb{P}_{H_0}(\overline{X}_n \geq x_2) \leq \alpha/2$
  - la $p$-valeur est $2 \times \min(\mathbb{P}_{H_0}\left(\overline{X}_n \geq \overline{X}_{n,\mathrm{obs}}\right), \mathbb{P}_{H_0}(\overline{X}_n \leq \overline{X}_{n,\mathrm{obs}}))$

  et l'on sait que $x_1$, $x_2$ sont de la forme

$$x_1 = m_0 - u \times \frac{\sigma}{\sqrt{n}} \qquad x_2 = m_0 + u \times \frac{\sigma}{\sqrt{n}}$$

# Chapitre 3 : exercices d'autoévaluation

Reprendre les calculs de seuils dans les cas des dix Norvégiens, des dix Boliviens et des dix Italiens pour un niveau $\alpha = 4\%$.

Les réponses sont à la page suivante.

# Chapitre 3 : exercices d'autoévaluation

- **Norvégiens :** on reprend à la page 108 : on cherche le plus petit $x_1'$ tel que $\mathbb{P}(\overline{X}_{10} \geq x_1') \leq 4\%$ si $\overline{X}_{10} \sim \mathcal{N}(179; \frac{7^2}{10})$. En appliquant exactement les mêmes calculs qu'en page 109 (faites les vous-mêmes !) on voit qu'on cherche en fait le plus petit $x_1'$ tel que $\mathbb{P}(U \leq \frac{x_1'-179}{7/\sqrt{10}}) \geq 1 - 0{,}04$ où $U \sim \mathcal{N}(0;1)$. La table donne $\frac{x_1'-179}{7/\sqrt{10}} = 1{,}76$ d'où $x_1' = 179 + 1{,}76 \times \frac{7}{\sqrt{10}} \simeq 182{,}86$.
- **Boliviens :** on reprend à la page 112 : on cherche le plus grand $x_2'$ tel que $\mathbb{P}(\overline{X}_{10} \geq x_2') \leq 4\%$ si $\overline{X}_{10} \sim \mathcal{N}(179; \frac{7^2}{10})$. En appliquant exactement les mêmes calculs qu'en page 113 (faites les vous-mêmes !) on voit qu'on cherche en fait le plus petit $-\frac{x_2'-179}{7/\sqrt{10}}$ tel que $\mathbb{P}(U \leq -\frac{x_2'-179}{7/\sqrt{10}}) \geq 1 - 0{,}04$ où $U \sim \mathcal{N}(0;1)$. La table donne $-\frac{x_2'-179}{7/\sqrt{10}} = 1{,}76$ d'où $x_1' = 179 - 1{,}76 \times \frac{7}{\sqrt{10}} \simeq 175{,}10$.
- **Italiens :** on reprend page 116 : on cherche le plus grand $x_3'$ et le plus petit $x_4'$ tels que $\mathbb{P}(\overline{X}_{10} \leq x_3') \leq 2\%$ et $\mathbb{P}(\overline{X}_{10} \geq x_4') \leq 2\%$ si $\overline{X}_{10} \sim \mathcal{N}(179; \frac{7^2}{10})$. Cherchons $x_4'$ : en appliquant exactement les mêmes calculs que dans le cas des Norvégiens, on voit qu'on cherche le plus petit $x_4'$ tel que $\mathbb{P}(U \leq \frac{x_4'-179}{7/\sqrt{10}}) \geq 1 - 0{,}02$ où $U \sim \mathcal{N}(0;1)$. La table donne $\frac{x_4'-179}{7/\sqrt{10}} = 2{,}06$ d'où $x_4' = 179 + 2{,}06 \times \frac{7}{\sqrt{10}}$. On sait qu'alors $x_4' = 179 - 2{,}06 \times \frac{7}{\sqrt{10}}$. On calcule $x_3' = 174{,}44$ et $x_4' = 183{,}56$.

# Tests sur une population normale (de variance inconnue)

# Pour quoi faire ?

Reprenons l'exemple 2 :

Vous arrivez en Norvège, et les dix premiers hommes (jeunes adultes) que vous rencontrez mesurent 180, 185, 184, 188, 182, 187, 182, 185, 181, 186 cm. Sachant que la taille moyenne des Français de la même catégorie est de 179 cm, doit-on en conclure que les Norvégiens sont en moyenne plus grands que les Français ?

On sait le traiter quand on connaît la variance de la taille chez les Français et en supposant que c'est la même chez les Norvégiens.

# Pour quoi faire ?

Comment le traiter

- quand on connaît la variance de la taille chez les Français mais sans supposer que c'est la même chez les Norvégiens ?
- sans connaître les variances de la taille, ni chez les Norvégiens, ni chez les Français ?

Il sera naturel :

- quand on connaît la variance de la taille chez les Français, de tester l'hypothèse que c'est la même chez les Norvégiens ;
- si la réponse est non, ou si l'on ne connaît pas la variance de la taille chez les Français, de tester malgré tout l'égalité des moyennes.

C'est ce que l'on va apprendre à faire dans ce chapitre.

# Exemples

On va traiter les deux questions suivantes (le premier est le véritable exemple 2, le deuxième est une variante).

2. Vous arrivez en Norvège, et les dix premiers hommes (jeunes adultes) que vous rencontrez mesurent 180, 185, 184, 188, 182, 187, 182, 185, 181, 186 cm. Sachant que la taille moyenne des Français de la même catégorie est de 179, doit-on en conclure que les Norvégiens sont en moyenne plus grands que les Français ?

   (on ne suppose connue aucune des deux variances)

2'. Vous arrivez en Norvège, et les dix premiers hommes (jeunes adultes) que vous rencontrez mesurent 180, 185, 184, 188, 182, 187, 182, 185, 181, 186 cm. Sachant que la taille des Français de la même catégorie a une variance de $7^2$, doit-on en conclure que la variance de la taille des Norvégiens est la même que celle des Français ?

# Deux nouvelles familles de lois

Pour répondre à ces deux questions, il va nous falloir introduire deux nouvelles lois de probabilité

- pour répondre à la deuxième : les lois du $\chi^2$,
- pour répondre à la première : les lois de Student.

(et nous allons le faire dans cet ordre)

Nous commençons donc avec les lois du $\chi^2$.

# Lois du $\chi^2$ : définition

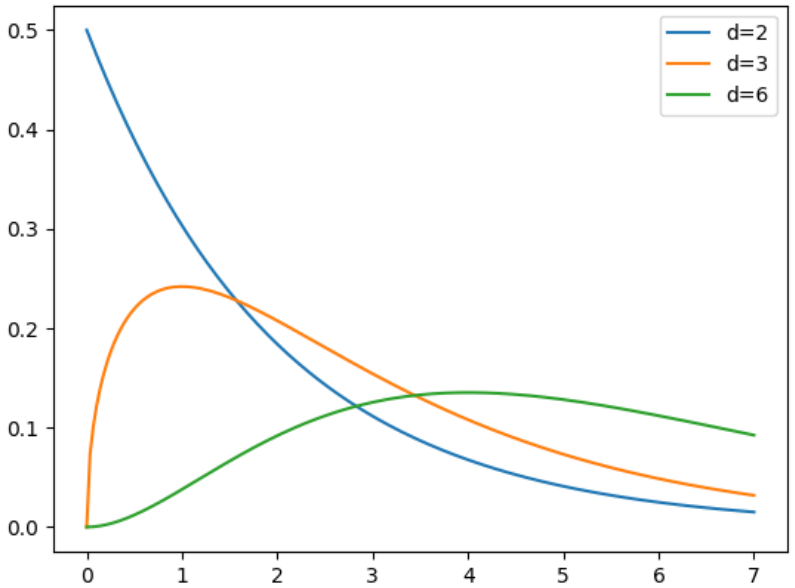
Si $U_1, \ldots, U_d$ sont $d$ variables indépendantes de même loi $\mathcal{N}(0,1)$, la variable

$$U_1^2 + \ldots + U_d^2$$

suit une *loi du $\chi^2$ à d degrés de liberté*, notée $\chi^2(d)$

Une telle variable aléatoire a les propriétés suivantes :

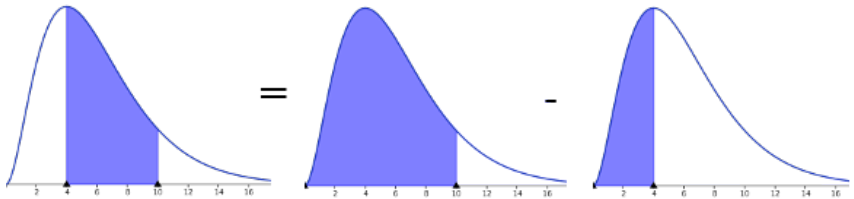
- c'est une loi à densité
  (que l'on sait exprimer mais qui ne nous servira pas)
- elle ne prend que des valeurs positives.



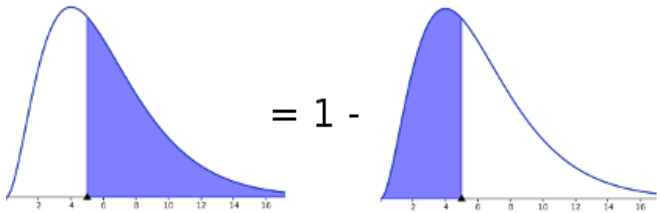
# Lois du $\chi^2$ : propriétés

si $Z \sim \chi^2(d)$ alors

$$\mathbb{P}(a \leq Z \leq b) = \mathbb{P}(Z \leq b) - \mathbb{P}(Z \leq a)$$




$$\mathbb{P}(Z \geq b) = 1 - \mathbb{P}(Z \leq b)$$




Attention, il n'y a pas de propriétés de symétrie comme pour la normale

# Lois du $\chi^2$ : tables de valeurs numériques

Les tables des lois du $\chi^2$ donnent pour un $d$ et un $\alpha$ donnés, le $t$ tel que si $Z \sim \chi^2(d)$

$$\mathbb{P}(Z \leq t) = \alpha$$

mais dans la table, $d$ est noté $n$ et $\alpha$ est noté $p$

Par exemple :

le $t$ tel que $\mathbb{P}(Z \leq t) = 10\%$ si $Z \sim \chi^2(5)$
est $t = 1{,}6103$

| $n \setminus p$ | 0.025 | 0.050 | 0.1 |
|---|---|---|---|
| 1 | 0.0010 | 0.0039 | 0.0158 |
| 2 | 0.0506 | 0.1026 | 0.2107 |
| 3 | 0.2158 | 0.3518 | 0.5844 |
| 4 | 0.4844 | 0.7107 | 1.0636 |
| 5 | 0.8312 | 1.1455 | 1.6103 |
| 6 | 1.2373 | 1.6354 | 2.2041 |
| 7 | 1.6899 | 2.1673 | 2.8331 |
| 8 | 2.1797 | 2.7326 | 3.4895 |

le $t$ tel que $\mathbb{P}(Z \leq t) = 98\%$ si $Z \sim \chi^2(7)$
est $t = 16{,}6224$

| $n \setminus p$ | 0.975 | 0.980 | 0.990 |
|---|---|---|---|
| 1 | 5.0239 | 5.4119 | 6.6349 |
| 2 | 7.3778 | 7.8240 | 9.2103 |
| 3 | 9.3484 | 9.8374 | 11.3449 |
| 4 | 11.1433 | 11.6678 | 13.2767 |
| 5 | 12.8325 | 13.3882 | 15.0863 |
| 6 | 14.4494 | 15.0332 | 16.8119 |
| 7 | 16.0128 | 16.6224 | 18.4753 |
| 8 | 17.5345 | 18.1682 | 20.0902 |

# Lois du $\chi^2$ : tables de valeurs numériques

Attention, il y a un tableau pour $p \leq 20\%$ et un autre pour $p \geq 80\%$

Le $t$ tel que $\mathbb{P}(Z \leq t) = \alpha$ est *unique* et c'est aussi

- le plus grand $t$ tel que $\mathbb{P}(Z \leq t) \leq \alpha$
- le plus petit $t$ tel que $\mathbb{P}(Z \leq t) \geq \alpha$

Cette présentation (différente des tables précédentes) fait que dans la suite :

- il sera particulièrement facile de trouver les seuils,
- il sera un peu plus compliqué de trouver les $p$-valeurs (voir un exemple page 143).

# Lois du $\chi^2$ : application

Si $X_1, \ldots, X_n$ est un $n$-échantillonde loi $\mathcal{N}(m; \sigma^2)$ alors

$$\frac{(X_1 - m)^2}{\sigma^2} + \ldots + \frac{(X_n - m)^2}{\sigma^2} \text{ suit une loi } \chi^2(n)$$

car chaque $\frac{X_i - m}{\sigma}$ suit la loi $\mathcal{N}(0,1)$ et les $\frac{X_i - m}{\sigma}$ sont indépendants

Ceci nous permet de faire des tests sur $\sigma$ si $m$ est connu !

# Lois du $\chi^2$ : application

par exemple pour tester $H_0 : \sigma = \sigma_0$ contre $H_1 : \sigma > \sigma_0$, on peut utiliser

$$Z = \frac{(X_1 - m)^2}{{\sigma_0}^2} + \ldots + \frac{(X_n - m)^2}{{\sigma_0}^2}$$

comme variable de test car

- sous l'hypothèse $H_0$, elle suit la loi $\chi^2(n)$,
- sous l'hypothèse $H_1$, $Z$ est une variable de loi $\chi^2(n)$ multipliée par un coefficient $> 1$,

donc $Z$ est plutôt plus grand sous $H_1$ que sous $H_0$.

pour préciser ce que l'on dit ci-dessus sur le cas $H_1$ :

$$Z = \frac{\sigma^2}{\sigma_0^2} \times \left( \frac{(X_1 - m)^2}{\sigma^2} + \ldots + \frac{(X_n - m)^2}{\sigma^2} \right)$$

# Lois du $\chi^2$ : application

On peut donc tester $H_0 : \sigma = \sigma_0$ contre $H_1 : \sigma > \sigma_0$ :

- en utilisant la variable de test $Z$
- avec une zone de rejet de la forme $[Z \geq z_0]$
- où $z_0$ est le plus petit possible tel que $\mathbb{P}_{H_0}(Z \geq z_0) \leq \alpha$
- que l'on peut déterminer en utilisant le fait que $Z \sim \chi^2(n)$ sous $H_0$

On adapte facilement le raisonnement aux cas

- $H_0 : \sigma = \sigma_0$ contre $H_1 : \sigma < \sigma_0$,
- $H_0 : \sigma = \sigma_0$ contre $H_1 : \sigma \neq \sigma_0$

Attention, on a besoin de connaître $m$ !

# Tests sur la variance d'une population normale à moyenne connue : résumé

Si l'on veut tester au niveau $\alpha$ la variance $\sigma^2$ d'une d'un $n$-échantillon $X_1, \ldots, X_n$ de $\mathcal{N}(m; \sigma^2)$ où $m$ est connu, dans tous les cas on utilise la variable

$$Z = \frac{(X_1 - m)^2}{\sigma_0^2} + \ldots + \frac{(X_n - m)^2}{\sigma_0^2}$$

il y a trois possibilités :

- pour un test $H_0 : \sigma = \sigma_0$ contre $H_1 : \sigma > \sigma_0$ (où $\sigma_0$ est connu)
  - $Z$ sera plutôt plus grand sous $H_1$ que sous $H_0$, donc
  - on rejette $H_0$ si $Z$ est "trop grand" : la zone de rejet est de la forme $[Z \geq z_0]$
  - le $z_0$ cherché est le plus petit réel $z$ tel que $\mathbb{P}_{H_0}(Z \geq z) \leq \alpha$
  - la $p$-valeur est $\mathbb{P}_{H_0}(Z \geq Z_{\text{obs}})$

# Tests sur la variance d'une population normale à moyenne connue : résumé

- pour un test $H_0 : \sigma = \sigma_0$ contre $H_1 : \sigma < \sigma_0$ (où $\sigma_0$ est connu)
  - $Z$ sera plutôt plus petit sous $H_1$ que sous $H_0$, donc
  - on rejette $H_0$ si $Z$ est "trop petit" : la zone de rejet est de la forme $[Z \leq z_0]$
  - le $z_0$ cherché est le plus grand $z$ tel que $\mathbb{P}_{H_0}(Z \leq z) \leq \alpha$
  - la $p$-valeur est $\mathbb{P}_{H_0}(Z \leq Z_{\text{obs}})$
- pour un test $H_0 : \sigma = \sigma_0$ contre $H_1 : \sigma \neq \sigma_0$ (où $\sigma_0$ est connu)
  - $Z$ sera plutôt plus soit plus petit, soit plus grand sous $H_1$ que sous $H_0$, donc
  - on rejette $H_0$ si $Z$ est "trop petit ou trop grand" : la zone de rejet est de la forme $[Z \leq z_1$ ou $Z \geq z_2]$
  - $z_1$ est le plus grand réel $z$ tel que $\mathbb{P}_{H_0}(Z \leq z) \leq \alpha/2$
  - $z_2$ est le plus petit réel tel que $\mathbb{P}_{H_0}(Z \geq z_2) \leq \alpha/2$
  - la $p$-valeur est $2 \times \min(\mathbb{P}_{H_0}(Z \geq Z_{\text{obs}}), \mathbb{P}_{H_0}(Z \leq Z_{\text{obs}}))$

Dans tous les cas on utilise le fait que sous $H_0$, $Z$ suit une loi $\chi^2(n)$

# Lois du $\chi^2$ : application

Et si $m$ est inconnu ?

**Propriété :** si $X_1, \ldots, X_n$ est un $n$-échantillon de loi $\mathcal{N}(m; \sigma^2)$ alors

$$\frac{(X_1 - \overline{X}_n)^2}{\sigma^2} + \ldots + \frac{(X_n - \overline{X}_n)^2}{\sigma^2} \text{ suit une loi } \chi^2(n-1)$$

(on ne démontrera pas ce résultat)

Ceci nous permet de faire des tests sur $\sigma$ si $m$ est inconnu !

# Lois du $\chi^2$ : application

par exemple pour tester $H_0 : \sigma = \sigma_0$ contre $H_1 : \sigma > \sigma_0$, on peut utiliser

$$Z' = \frac{(X_1 - \overline{X}_n)^2}{{\sigma_0}^2} + \ldots + \frac{(X_n - \overline{X}_n)^2}{{\sigma_0}^2}$$

comme variable de test car

- sous l'hypothèse $H_0$, elle suit une loi $\chi^2(n-1)$,
- sous l'hypothèse $H_1$, $Z'$ est une variable de loi $\chi^2(n-1)$ multipliée par un coefficient $> 1$,

donc $Z'$ est plutôt plus grand sous $H_1$ que sous $H_0$.

pour préciser ce que l'on dit ci-dessus sur le cas $H_1$ :

$$Z' = \frac{\sigma^2}{\sigma_0^2} \times \left(\frac{(X_1 - \overline{X}_n)^2}{\sigma^2} + \ldots + \frac{(X_n - \overline{X}_n)^2}{\sigma^2}\right)$$

# Lois du $\chi^2$ : application

On peut donc tester $H_0 : \sigma = \sigma_0$ contre $H_1 : \sigma > \sigma_0$ :

- en utilisant la variable de test $Z'$
- avec une zone de rejet de la forme $[Z' \geq z_0]$
- où $z_0$ est le plus petit possible tel que $\mathbb{P}_{H_0}(Z' \geq z_0) \leq \alpha$
- que l'on peut déterminer en utilisant le fait que $Z' \sim \chi^2(n-1)$ sous $H_0$

Encore une fois, on adapte facilement le raisonnement aux cas

- $H_0 : \sigma = \sigma_0$ contre $H_1 : \sigma < \sigma_0$,
- $H_0 : \sigma = \sigma_0$ contre $H_1 : \sigma \neq \sigma_0$

# Tests sur la variance d'une population normale à moyenne inconnue

On va traiter l'exemple suivant :

Vous arrivez en Norvège, et les dix premiers hommes (jeunes adultes) que vous rencontrez mesurent 180, 185, 184, 188, 182, 187, 182, 185, 181, 186 cm. Sachant que la taille des Français de la même catégorie a une variance de $7^2$, doit-on en conclure que la variance de la taille des Norvégiens est la même que celle des Français ?

(on traitera le cas où la moyenne de la taille des Norvégiens est inconnue, en testant $H_0 : \sigma = 7$ contre $H_1 : \sigma < 7$, et on fera le test au niveau $\alpha = 2\%$)

# Tests sur la variance d'une population normale à moyenne inconnue

On se base sur l'analyse donnée en pages 139 et 140, mais on écrit directement le tout en sept points.

1. **le modèle** : les tailles $X_1, \ldots, X_{10}$ des dix Norvégiens observés forment un 10-échantillon de loi $\mathcal{N}(m; \sigma^2)$.
2. **les hypothèses** : $H_0 : \sigma = 7$ et $H_1 : \sigma < 7$.
3. **la variable de test** : on utilise la variable
$$Z' = \frac{(X_1 - \overline{X}_n)^2}{7^2} + \ldots + \frac{(X_n - \overline{X}_n)^2}{7^2}.$$
4. **la forme de la zone de rejet** : sous l'hypothèse $H_0$, $Z'$ suit une loi $\chi^2(9)$ et sous $H_1$, $Z'$ est $\sigma^2/7^2$ fois une variable de loi $\chi^2(9)$, donc $Z'$ est plutôt plus petite sous l'hypothèse $H_1$ que sous l'hypothèse $H_0$ et on va rejeter $H_0$ si $Z'$ est "trop petit". La zone de rejet est donc de la forme $[Z' \leq z_0]$.
5. **le seuil de la zone de rejet** : au niveau $\alpha = 2\%$, on cherche le plus grand $z_0$ tel que $\mathbb{P}(Z' \leq z_0) \leq 2\%$ si $Z' \sim \chi^2(9)$. La table donne $z_0 = 2{,}5324$.

# Tests sur la variance d'une population normale à moyenne inconnue

6. **la conclusion** : pour calculer la valeur observée de $Z'$, il faut d'abord calculer celle de $\overline{X}_{10}$ :

$$\overline{X}_{10,\mathrm{obs}} = \frac{180+185+184+188+182+187+182+185+181+186}{10} = 184$$

$$Z' = \frac{(180-184)^2}{7^2} + \frac{(185-184)^2}{7^2} + \ldots + \frac{(186-184)^2}{7^2} \simeq 1{,}3061$$

   On rejette donc l'hypothèse $H_0$ au niveau 2%.

7. **la *p*-valeur** : la zone de rejet étant de la forme $[Z' \leq z_0]$ et l'observation étant $Z' = 1{,}3061$ la *p*-valeur vaut $\mathbb{P}_{H_0}(Z' \leq 1{,}3061)$. On ne peut lire directement cette valeur dans la table ; en revanche on peut lire que $\mathbb{P}_{H_0}(Z' \leq 1{,}1519) = 0{,}001$ et $\mathbb{P}_{H_0}(Z' \leq 1{,}7349) = 0{,}005$ et on en déduit que la *p*-valeur $\mathbb{P}_{H_0}(Z' \geq 1{,}3061)$ est comprise entre 0,1% et 0,5%.

# Tests sur la variance d'une population normale à moyenne inconnue : résumé

Si l'on veut tester au niveau $\alpha$ la variance $\sigma^2$ d'un $n$-échantillon $X_1, \ldots, X_n$ de $\mathcal{N}(m; \sigma^2)$ où $m$ est inconnu, dans tous les cas on utilise la variable

$$Z' = \frac{(X_1 - \overline{X}_n)^2}{\sigma_0^2} + \ldots + \frac{(X_n - \overline{X}_n)^2}{\sigma_0^2}$$

il y a trois possibilités :

- pour un test $H_0 : \sigma = \sigma_0$ contre $H_1 : \sigma > \sigma_0$ (où $\sigma_0$ est connu)
  - $Z'$ sera plutôt plus grand sous $H_1$ que sous $H_0$, donc
  - on rejette $H_0$ si $Z'$ est "trop grand" : la zone de rejet est de la forme $[Z' \geq z_0]$
  - le $z_0$ cherché est le plus petit réel $z$ tel que $\mathbb{P}_{H_0}(Z' \geq z) \leq \alpha$
  - la $p$-valeur est $\mathbb{P}_{H_0}(Z' \geq Z'_{\text{obs}})$

# Tests sur la variance d'une population normale à moyenne inconnue : résumé

- pour un test $H_0 : \sigma = \sigma_0$ contre $H_1 : \sigma < \sigma_0$ (où $\sigma_0$ est connu)
  - $Z'$ sera plutôt plus petit sous $H_1$ que sous $H_0$, donc
  - on rejette $H_0$ si $Z'$ est "trop petit" : la zone de rejet est de la forme $[Z' \leq z_0]$
  - le $z_0$ cherché est le plus petit grand $z$ tel que $\mathbb{P}_{H_0}(Z' \leq z) \leq \alpha$
  - la $p$-valeur est $\mathbb{P}_{H_0}(Z' \leq Z'_{\text{obs}})$
- pour un test $H_0 : \sigma = \sigma_0$ contre $H_1 : \sigma \neq \sigma_0$ (où $\sigma_0$ est connu)
  - $Z'$ sera plutôt plus soit plus petit, soit plus grand sous $H_1$ que sous $H_0$, donc
  - on rejette $H_0$ si $Z'$ est "trop petit ou trop grand" : la zone de rejet est de la forme $[Z' \leq z'_1$ ou $Z' \geq z'_2]$
  - $z'_1$ est le plus grand réel $z$ tel que $\mathbb{P}_{H_0}(Z' \leq z) \leq \alpha/2$
  - $z'_2$ est le plus petit réel tel que $\mathbb{P}_{H_0}(Z' \geq z) \leq \alpha/2$
  - la $p$-valeur est $2 \times \min(\mathbb{P}_{H_0}(Z' \geq Z'_{\text{obs}}), \mathbb{P}_{H_0}(Z' \leq Z'_{\text{obs}}))$

Dans tous les cas on utilise le fait que sous $H_0$, $Z'$ suit une loi $\chi^2(n-1)$

# Tests sur la moyenne d'une population normale à variance inconnue

On va maintenant introduire les lois de Student qui nous permettront de traiter l'exemple 2.

# Lois de Student : définition

Si

- $U$ suit une loi normale centrée réduite $\mathcal{N}(0,1)$,
- $Z$ suit une loi du $\chi^2$ à $d$ degrés de liberté $\chi^2(d)$,
- et que $U$ et $Z$ sont indépendants,

alors la variable

$$T = \frac{U}{\sqrt{Z/d}}$$

suit une loi appelée loi de Student à $d$ degrés de liberté, notée $\mathcal{T}(d)$.

Une telle variable aléatoire admet une densité
(que l'on sait exprimer mais ne nous servira pas).

# Lois de Student : propriétés

- cette densité ressemble à celle d'une loi normale centrée réduite
- en fait pour $d \to \infty$ on sait par la loi des grands nombres que $Z/d \to 1$

  car $Z/d$ a la même loi que $\frac{U_1+\ldots+U_d}{d}$, qui tend vers $\mathbb{E}(U_1^2) = 1$ quand $d \to \infty$ par la loi des grands nombres

  donc $T \simeq U$, autrement dit une loi de Student avec $d$ très grand est presque une loi $\mathcal{N}(0,1)$.

# Lois de Student : tables de valeurs numériques

Les tables des lois de Student donnent pour $d$ et $\alpha$ donnés, le $t$ tel que si $T \sim \mathcal{T}(d)$

$$\mathbb{P}(T \leq t) = \alpha$$

mais dans la table, $d$ est noté $n$ et $\alpha$ est noté $p$

Le $t$ tel que $\mathbb{P}(T \leq t) = \alpha$ est *unique* et c'est aussi

- le plus grand $t$ tel que $\mathbb{P}(Z \leq t) \leq \alpha$
- le plus petit $t$ tel que $\mathbb{P}(Z \leq t) \geq \alpha$

Par exemple

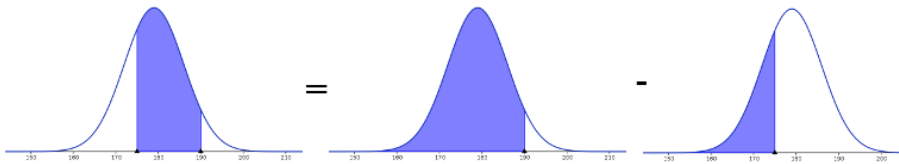
| $n \setminus p$ | 0.550 | 0.650 | 0.750 | 0.850 | 0.900 | 0.950 | 0.975 | 0.990 | 0.995 | 0.999 | 0.995 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.158 | 0.510 | 1.000 | 1.963 | 3.078 | 6.314 | 12.706 | 31.821 | 63.657 | 318.309 | 636.619 |
| 2 | 0.142 | 0.445 | 0.816 | 1.386 | 1.886 | 2.920 | 4.303 | 6.965 | 9.925 | 22.327 | 31.599 |
| 3 | 0.137 | 0.424 | 0.765 | 1.250 | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 | 10.215 | 12.924 |
| 4 | 0.134 | 0.414 | 0.741 | 1.190 | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 | 7.173 | 8.610 |
| 5 | 0.132 | 0.408 | 0.727 | 1.156 | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 | 5.893 | 6.869 |
| 6 | 0.131 | 0.404 | 0.718 | 1.134 | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 | 5.208 | 5.959 |

indique que le $t$ tel que $\mathbb{P}(T \leq t) = 0{,}975$ pour $T \sim \mathcal{T}(5)$ est $t_0 = 2{,}571$

# Lois de Student : propriétés

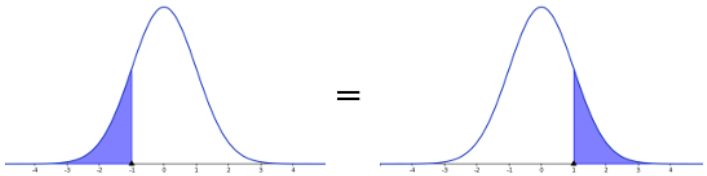
Les lois de Student suivent les mêmes règles de calcul que la loi normale, à savoir que

$$\mathbb{P}(a \leq T \leq b) = \mathbb{P}(T \leq b) - \mathbb{P}(T \leq a)$$



$$\mathbb{P}(T \geq b) = 1 - \mathbb{P}(T \leq b)$$

$$\mathbb{P}(T \leq -t) = \mathbb{P}(T \geq +t)$$



# Lois de Student : tables de valeurs numériques

Donnons des exemples de calcul : si l'on cherche

- le plus petit $t$ tel que $\mathbb{P}(T \geq t) \leq 0{,}05$ pour $T \sim \mathcal{T}(6)$ :

  comme $\mathbb{P}(T \geq t) = 1 - \mathbb{P}(T \leq t)$ cela revient à avoir $\mathbb{P}(T \leq t) \geq 0{,}95$. La table donne que la valeur pour laquelle on a l'égalité est $t_0 = 1{,}943$

- le plus grand $t$ tel que $\mathbb{P}(T \leq t) \leq 0{,}01$ pour $T \sim \mathcal{T}(4)$ :

  on n'a pas $\alpha = 0{,}01$ dans la table mais $\mathbb{P}(T \leq t) = \mathbb{P}(T \geq -t)$ et c'est aussi $1 - \mathbb{P}(T \leq -t)$ donc c'est équivalent à avoir $\mathbb{P}(T \leq -t) \geq 0{,}99$. La table donne que la valeur pour laquelle on a l'égalité est $t_0 = -3{,}747$

- le plus grand $t$ tel que $\mathbb{P}(-t \leq T \leq +t) \leq 0{,}99$ pour $T \sim \mathcal{T}(5)$ :

  on a $\mathbb{P}(-t \leq T \leq +t) = \mathbb{P}(T \leq +t) - \mathbb{P}(T \leq -t)$ et $\mathbb{P}(T \leq -t) = \mathbb{P}(T \geq +t) = 1 - \mathbb{P}(T \leq t)$. On cherche donc $t$ tel que $2\mathbb{P}(T \leq t) - 1 \leq 0{,}99$ donc $\mathbb{P}(T \leq t) \leq 0{,}995$. La table donne que la valeur pour laquelle on a l'égalité est $t_0 = 4{,}032$.

# Lois de Student : application

Si $X_1, \ldots, X_n$ est un $n$-échantillon de loi $\mathcal{N}(m; \sigma^2)$ alors

$$\overline{X}_n = \frac{X_1 + \ldots + X_n}{n} \sim \mathcal{N}\left(m, \frac{\sigma^2}{n}\right)$$

donc

$$\sqrt{n}\frac{\overline{X}_n - m}{\sigma} \sim \mathcal{N}(0,1)$$

$$S_n^2 \overset{\text{défi}}{=} \frac{1}{n-1}\left((X_1 - \overline{X}_n)^2 + \ldots + (X_n - \overline{X}_n)^2\right)$$

vérifie que

$$(n-1)\frac{S_n^2}{\sigma^2} \sim \chi^2(n-1)$$

et on a

**Propriété :** $\overline{X}_n$ et $S_n^2$ sont indépendants

(on ne démontrera pas ce résultat)

# Lois de Student : application

Après un peu de simplification, on voit que

$$\sqrt{n}\,\frac{\overline{X}_n - m}{S_n} \sim \mathcal{T}(n-1).$$

Ceci nous permet de faire des tests sur $m$ si $\sigma$ est inconnu !

# Lois de Student : application

par exemple pour tester $H_0 : m = m_0$ contre $H_1 : m > m_0$, on peut utiliser la variable de test

$$T = \sqrt{n}\,\frac{\overline{X}_n - m_0}{S_n}$$

comme variable de test car

- sous l'hypothèse $H_0$ elle suit la loi $\mathcal{T}(n-1)$,
- sous l'hypothèse $H_1$, $T$ est une variable de loi $\mathcal{T}(n-1)$ *plus un terme positif*, donc $T$ est plutôt plus grand sous $H_1$ que sous $H_0$.

pour préciser ce que l'on dit ci-dessus sur le cas $H_1$ :

$$T = \sqrt{n}\,\frac{\overline{X}_n - m}{S_n} + \sqrt{n}\,\frac{m - m_0}{S_n}$$

et le dernier terme est strictement positif si $m > m_0$.

# Lois de Student : application

On peut donc tester $H_0 : m = m_0$ contre $H_1 : m > m_0$ :

- en utilisant la variable de test $T$
- avec une zone de rejet de la forme $[T \geq t_0]$
- où $t_0$ est le plus petit possible tel que $\mathbb{P}_{H_0}(T \geq t_0) \leq \alpha$
- que l'on peut déterminer en utilisant le fait que $T \sim \mathcal{T}(n-1)$ sous $H_0$

Encore une fois, on adapte facilement le raisonnement aux cas

- $H_0 : m = m_0$ contre $H_1 : m < m_0$,
- $H_0 : m = m_0$ contre $H_1 : m \neq m_0$

# Tests sur la moyenne d'une population normale à variance inconnue

On va (enfin) traiter l'exemple suivant :

Vous arrivez en Norvège, et les dix premiers hommes (jeunes adultes) que vous rencontrez mesurent 180, 185, 184, 188, 182, 187, 182, 185, 181, 186 cm. Sachant que la taille moyenne des Français de la même catégorie est de 179, doit-on en conclure que les Norvégiens sont en moyenne plus grands que les Français ?

(on fera le test au niveau $\alpha = 3\%$ ; on ne suppose connue aucune des deux variances)

# Tests sur la moyenne d'une population normale à variance inconnue

On se base sur l'analyse donnée en page 154, mais on écrit directement le tout en sept points.

1. **le modèle** : les tailles $X_1, \ldots, X_{10}$ des dix Norvégiens observés forment un 10-échantillon de loi $\mathcal{N}(m; \sigma^2)$.
2. **les hypothèses** : $H_0$ : $m = 179$ et $H_1$ : $m > 179$.
3. **la variable de test** : on utilise la variable
$$T = \sqrt{10}\,\frac{\overline{X}_{10} - 179}{S_{10}} \text{ où } \overline{X}_{10} = \frac{X_1 + \ldots + X_{10}}{10}.$$
4. **la forme de la zone de rejet** : on a $T = \sqrt{10}\,\frac{\overline{X}_{10} - m}{S_{10}} + \sqrt{10}\,\frac{m - 179}{S_{10}}$. Le premier terme suit une loi $\mathcal{T}(9)$ ; le deuxième est nul si $H_0$ est vraie et strictement positif si $H_1$ est vraie. Par conséquent, $T$ est plutôt plus grand sous $H_1$ que sous $H_0$, on va rejeter $H_0$ si $T$ est "trop grand". La zone de rejet est donc de la forme $[T \geq t_0]$.
5. **le seuil de la zone de rejet** : au niveau $\alpha = 3\%$, on cherche le plus petit $t_0$ tel que $\mathbb{P}(T \geq t_0) \leq 3\%$ si $T \sim \mathcal{T}(9)$. Il y a un peu de calcul.

# Tests sur la moyenne d'une population normale à variance inconnue

On cherche le plus petit $t_0$ tel que $\mathbb{P}(T \geq t_0) \leq 3\%$ si $T \sim \mathcal{T}(9)$

on applique la deuxième règle page 150 :

$$\mathbb{P}(T \geq t_0) = 1 - \mathbb{P}(T \leq t_0)$$

on cherche donc le plus petit $t_0$ tel que

$$\mathbb{P}(T \leq t_0) \geq 1 - 0{,}03.$$

Dans la table, on peut seulement lire que, si $T \sim \mathcal{T}(9)$, alors $\mathbb{P}(T \leq 1{,}833) = 0{,}95$ et $\mathbb{P}(T \leq 2{,}262) = 0{,}975$. Par conséquent $t_0$ est compris entre 1,833 et 2,262.

# Tests sur la moyenne d'une population normale à variance inconnue

6. **la conclusion** : pour calculer la valeur observée de $T$, il faut d'abord calculer celles de $\overline{X}_{10}$ et $S_{10}$ :

$$\overline{X}_{10,\text{obs}} = \frac{180+185+184+188+182+187+182+185+181+186}{10} = 184$$
$$S^2_{10,\text{obs}} = \frac{(180-184)^2+(185-184)^2+\ldots+(186-184)^2}{10-1} \simeq 7{,}11$$
$$T_{\text{obs}} = \sqrt{10}\,\frac{\overline{X}_{10,\text{obs}}-179}{S_{10,\text{obs}}} = \sqrt{10}\,\frac{184-179}{\sqrt{7{,}11}} \simeq 5{,}93$$

   on observe $T_{\text{obs}} = 5{,}93$ qui est supérieur à 1,833 donc on rejette $H_0$ au niveau 3%.

7. **la *p*-valeur** : la zone de rejet étant de la forme $[T \geq t_0]$ et l'observation étant $T_{\text{obs}} = 0{,}313$ la *p*-valeur vaut $\mathbb{P}_{H_0}(T \geq 5{,}93)$. Encore un coup il y a un peu de calculs. On applique la deuxième règle page 150 :

$$\mathbb{P}_{H_0}(T \geq 5{,}93) = 1 - \mathbb{P}_{H_0}(T \leq 5{,}93).$$

   Tout ce qu'on peut lire dans la table, c'est que $\mathbb{P}_{H_0}(T \leq 4{,}781) = 99{,}5\%$. Par conséquent $\mathbb{P}_{H_0}(T \leq 5{,}93)$ est supérieur à 99,5% et la *p*-valeur est inférieure à 0,05%.

# Comment se souvenir de l'expression de $T$ ?

Le test sur la moyenne d'une population normale à variance inconnue utilise la variable $T = \sqrt{n}\frac{\overline{X}_n - m_0}{S_n}$ qui suit la loi $\mathcal{T}(n-1)$ si $m = m_0$

comment se souvenir de son expression ?

# Comment se souvenir de l'expression de $T$ ?

on sait que dans le cas où $X_1, \ldots, X_n$ est un $n$-échantillon de loi $\mathcal{N}(m; \sigma^2)$, la moyenne empirique $\overline{X}_n$ suit une loi $\mathcal{N}(m; \frac{\sigma^2}{n})$. De manière équivalente,

$$\sqrt{n}\frac{\overline{X}_n - m}{\sigma} \text{ suit une loi } \mathcal{N}(0; 1).$$

Cependant, quand $\sigma$ est inconnu, on ne peut pas calculer cette variable de test à partir des valeurs observées de $X_1, \ldots, X_n$.

On le remplace donc par $S_n$ qui est un *estimateur* de $\sigma$.

C'est-à-dire ?

# Un mot sur les estimateurs

On a vu plusieurs exemples d'estimateurs :

La loi des grands nombres permet de montrer que si $X_1, \ldots, X_n$ est un $n$-échantillon *de n'importe quelle loi*

$$\overline{X}_n \overset{\text{défi}}{=} \frac{X_1+\ldots+X_n}{n} \underset{n\to\infty}{\longrightarrow} \mathbb{E}(X_1)$$

$$V_n^2 \overset{\text{défi}}{=} \frac{(X_1-\mathbb{E}(X_1))^2+\ldots+(X_n-\mathbb{E}(X_n))^2}{n} \underset{n\to\infty}{\longrightarrow} \mathbb{E}\big(X_1-\mathbb{E}(X_1)\big)^2 = \mathbb{V}(X_1)$$

Cela demande à peine plus de travail de montrer que

$$S_n^2 \overset{\text{défi}}{=} \frac{(X_1-\overline{X}_n)^2+\ldots+(X_n-\overline{X}_n)^2}{n-1} \underset{n\to\infty}{\longrightarrow} \mathbb{E}\big(X_1-\mathbb{E}(X_1)\big)^2 = \mathbb{V}(X_1)$$

Autrement dit, pour $n$ assez grand, $\overline{X}_n \simeq \mathbb{E}(X_1)$ et $V_n^2 \simeq S_n^2 \simeq \mathbb{V}(X_1)$.

Les variables $\overline{X}_n$, $V_n^2$ et $S_n^2$ sont appelées pour cette raisons des estimateurs de $\mathbb{E}(X_1)$ et $\mathbb{V}(X_1)$ respectivement.

# Comment se souvenir de l'expression de $T$ ?

0n sait que dans le cas où $X_1, \ldots, X_n$ est un $n$-échantillon de loi $\mathcal{N}(m; \sigma^2)$, la moyenne empirique $\overline{X}_n$ suit une loi $\mathcal{N}(m; \frac{\sigma^2}{n})$. De manière équivalente,

$$\sqrt{n}\frac{\overline{X}_n - m}{\sigma} \text{ suit une loi } \mathcal{N}(0; 1).$$

Cependant, quand $\sigma$ est inconnu, on ne peut pas calculer cette variable de test à partir des valeurs observées de $X_1, \ldots, X_n$.

On le remplace donc par $S_n$ qui est un *estimateur* de $\sigma$. La loi n'est plus alors une $\mathcal{N}(0; 1)$ mais une loi qui y ressemble, à savoir $\mathcal{T}(n-1)$ :

$$\sqrt{n}\frac{\overline{X}_n - m}{S_n} \text{ suit une loi } \mathcal{T}(n-1).$$

# Tests sur la moyenne d'une population normale à variance inconnue : résumé

Si l'on veut tester au niveau $\alpha$ la moyenne $m$ d'une d'un $n$-échantillon $X_1, \ldots, X_n$ de $\mathcal{N}(m; \sigma^2)$ où $\sigma$ est inconnu, dans tous les cas on utilise la variable

$$T = \sqrt{n}\,\frac{\overline{X}_n - m_0}{S_n}$$

il y a trois possibilités :

- pour un test $H_0 : m = m_0$ contre $H_1 : m < m_0$ (où $m_0$ est connu)
  - $T$ sera plutôt plus petit sous $H_1$ que sous $H_0$, donc
  - on rejette $H_0$ si $T$ est "trop petit" : la zone de rejet est de la forme $[T \leq t_0]$
  - le $t_0$ cherché est le plus grand $t$ tel que $\mathbb{P}_{H_0}(T \leq t) \leq \alpha$
  - la $p$-valeur est $\mathbb{P}_{H_0}(T \leq T_{\text{obs}})$
- pour un test $H_0 : m = m_0$ contre $H_1 : m > m_0$ (où $m_0$ est connu)
  - $T$ sera plutôt plus grand sous $H_1$ que sous $H_0$, donc
  - on rejette $H_0$ si $T$ est "trop grand" : la zone de rejet est de la forme $[T \geq t_0]$
  - le $t_0$ cherché est le plus petit $t$ tel que $\mathbb{P}_{H_0}(T \geq t) \leq \alpha$
  - la $p$-valeur est $\mathbb{P}_{H_0}(T \geq T_{\text{obs}})$

# Tests sur la moyenne d'une population normale à variance connue : résumé

- pour un test $H_0$ : $m = m_0$ contre $H_1$ : $m \neq m_0$ (où $m_0$ est connu)
  - $T$ sera plutôt plus soit plus petit, soit plus grand sous $H_1$ que sous $H_0$, donc
  - on rejette $H_0$ si $T$ est "trop petit ou trop grand" : la zone de rejet est de la forme $[T \leq t_1 \text{ ou } T \geq t_2]$
  - $t_1$ est le plus grand réel tel que $\mathbb{P}_{H_0}(T \leq t_1) \leq \alpha/2$
  - $t_2$ est le plus petit réel tel que $\mathbb{P}_{H_0}(T \geq t_2) \leq \alpha/2$
  - la $p$-valeur est $2 \times \min(\mathbb{P}_{H_0}(T \geq T_{\text{obs}}), \mathbb{P}_{H_0}(T \leq T_{\text{obs}}))$

  et la troisième règle page 148 montre que $t_1 = -t_2$ ; il suffit donc de chercher $t_2$.

# Récapitulatif des tests sur des populations normales

Si $X_1, \ldots, X_n$ est un $n$-échantillon de loi $\mathcal{N}(m; \sigma^2)$ et que l'on ne connaît pas à la fois $m$ et $\sigma^2$, il y a quatre possibilités de tests :

- on veut tester $m = m_0$ et on connaît $\sigma$ (cf. pages 120–121)
  on utilise la variable $\overline{X}_n$ qui suit une loi $\mathcal{N}(m_0, \sigma^2)$ si $m = m_0$
- on veut tester $m = m_0$ et on ne connaît pas $\sigma$ (cf. pages 164–165)
  on utilise la variable $\sqrt{n}\frac{\overline{X}_n - m_0}{S_n}$ qui suit une loi $\mathcal{T}(n-1)$ si $m = m_0$
- on veut tester $\sigma = \sigma_0$ et on connaît $m$ (cf. pages 136–137)
  on utilise la variable $nV_n^2/\sigma_0^2$ qui suit une loi $\chi^2(n)$ si $\sigma = \sigma_0$
- on veut tester $\sigma = \sigma_0$ et on ne connaît pas $m$ (cf. pages 144–145)
  on utilise la variable $(n-1)S_n^2/\sigma_0^2$ qui suit une loi $\chi^2(n-1)$ si $\sigma = \sigma_0$

# Chapitre 4 : exercices d'autoévaluation

1. Vous arrivez en Italie, et les dix premiers hommes (jeunes adultes) que vous rencontrez mesurent 173, 180, 176, 190, 185, 188, 164, 183, 173, 175 cm. Sachant que la taille des Français de la même catégorie a une variance de $7^2$, doit-on en conclure que la variance de la taille des Italiens est la même que celle des Français ? On fera le test au niveau 5%
2. Vous arrivez en Bolivie, et les dix premiers hommes (jeunes adultes) que vous rencontrez mesurent 160, 165, 164, 168, 162, 167, 162, 165, 161, 166 cm. Sachant que la taille moyenne des Français de la même catégorie est de 179 cm (et que la variance est inconnue) doit-on en conclure que les Boliviens sont en moyenne plus grands que les Français ?

# Chapitre 4 : exercices d'autoévaluation

1. Vous arrivez en Italie, et les dix premiers hommes (jeunes adultes) que vous rencontrez mesurent 173, 180, 176, 190, 185, 188, 164, 183, 173, 175 cm. Sachant que la taille des Français de la même catégorie a une variance de $7^2$, doit-on en conclure que la variance de la taille des Italiens est la même que celle des Français ? On fera le test au niveau 5%
2. Vous arrivez en Bolivie, et les dix premiers hommes (jeunes adultes) que vous rencontrez mesurent 160, 165, 164, 168, 162, 167, 162, 165, 161, 166 cm. Sachant que la taille moyenne des Français de la même catégorie est de 179 cm (et que la variance est inconnue) doit-on en conclure que les Boliviens sont en moyenne plus grands que les Français ?

# Tests du chi-deux d'adéquation et d'indépendance

# Pour quoi faire ?

Les résultats de cette section permettent de traiter l'exemple numéro 3 :

On examine 1000 Français originaires du Pays basque, et on constate que 740 ont pour groupe sanguin $O$, 250 un groupe sanguin $A$, 5 un groupe sanguin $B$ et 5 un groupe sanguin $AB$. Sachant que la répartition dans l'ensemble de la population française est 43%, 45%, 9%, 3%, doit-on en conclure que la population basque présente une particularité dans la population française ?

# Pour quoi faire ?

Les résultats de cette section permettent aussi de traiter l'exemple suivant :

On observe 1505 pommiers qui ont tous reçu un traitement parmi quatre possibles (notés $A$, $B$, $C$ et $D$) et on note s'ils ont été productifs (noté $P$) ou non productifs (noté $NP$). Les résultats sont résumés dans le tableau suivant :

| | $A$ | $B$ | $C$ | $D$ |
|---|---|---|---|---|
| $P$ | 156 | 113 | 128 | 185 |
| $NP$ | 203 | 266 | 258 | 196 |

Doit-on conclure que le traitement a un lien avec la productivité ?

# Pour quoi faire ?

un test comme celui sur les Basques est appelé test d'adéquation

(on se demande si la répartition des groupes sanguins observées chez les 1000 Basques est en adéquation avec une répartition connue par ailleurs)

un test comme celui sur les pommiers est appelé test d'indépendance

(on se demande si traitement et productivité sont indépendants d'après les observations effectuées sur 1505 pommiers)

La méthode pratique est très proche pour ces deux tests.

# Quelle variable de test ?

Un test est basé sur le choix d'une variable de test :

- dont on connaît la loi sous $H_0$,
- qui a un comportement différent sous $H_1$ et sous $H_0$.

Ici, ce sont des théorèmes mathématiques qui vont nous donner :

- la variable de test à choisir,
- sa loi sous $H_0$,
- son comportement sous $H_1$.

On va commencer par présenter le cas du test du chi-deux d'indépendance.

# Test du chi-deux d'adéquation : introduction

Supposons que l'on a $X_1, \ldots, X_n$ un $n$-échantillon de variables aléatoires qui peuvent prendre les valeurs $a_1, \ldots, a_d$.

**Exemple :** $X_1, \ldots, X_{1000}$ sont les groupes sanguins de 1000 Français originaires du Pays basque, qui valent donc O, A, B ou AB.

(on a donc $n = 1000$, $d = 4$ et les valeurs possibles sont 0, A, B et AB)

# Test du chi-deux d'adéquation : introduction

On veut tester si l'échantillon suit une loi que l'on nous propose.

Autrement dit, on nous propose des valeurs $p_1, \ldots, p_d$ et on veut voir si les observations sont cohérentes avec la proposition

$$\mathbb{P}(X = a_1) = p_1; \ldots; \mathbb{P}(X = a_d) = p_d.$$

**Exemple :** on veut savoir si la répartition des groupes sanguins dans la population basque (que l'on ne connaît pas) est la même que la répartition dans l'ensemble de la population française (que l'on connaît), qui est

$$p_{\mathrm{O}} = 0{,}43;\ p_{\mathrm{A}} = 0{,}45;\ p_{\mathrm{B}} = 0{,}09;\ p_{\mathrm{AB}} = 0{,}03.$$

# Test du chi-deux d'adéquation : notations

On note $N_k$ le nombre de fois où la valeur $a_k$ est observée parmi $X_1, \ldots, X_n$ (mais on utilisera souvent une notation plus explicite)

**Exemple :** On notera

$N_O$ le nb de personnes qui ont un gpe sanguin O,

$N_A$ le nb de personnes qui ont un gpe sanguin A,

$N_B$ le nb de personnes qui ont un gpe sanguin B,

$N_{AB}$ le nb de personnes qui ont un gpe sanguin AB.

(sur les 1000 personnes observées)

On a forcément

$$N_1 + \ldots + N_d = n$$

**Exemple :** on a observé $N_O = 740$, $N_A = 250$, $N_B = N_{AB} = 5$ et $740 + 250 + 5 + 5 = 1000$.

# Test du chi-deux d'adéquation : théorème

Le résultat théorique qui fait fonctionner le test d'adéquation est le suivant :

**Théorème**

si l'on note

$$Z = \frac{(N_1 - n\,p_1)^2}{n\,p_1} + \ldots + \frac{(N_d - n\,p_d)^2}{n\,p_d}$$

alors

- si l'on a bien $\mathbb{P}(X = x_k) = p_k$ pour tout $k = 1, \ldots, d$, alors pour $n$ grand, $Z$ suit (à peu près) une loi du chi-deux à $d - 1$ degrés de liberté,
- si ces relations ne sont pas toutes vraies, alors $Z \to \infty$ quand $n \to \infty$, donc pour $n$ grand $Z$ est "très grand".

Le **critère pratique** pour "$n$ grand" est que $n\,p_k \geq 5$ pour tous les $k = 1, \ldots, d$.

# Test du chi-deux d'adéquation : théorème

Autrement dit, *si* $n p_k \geq 5$ *pour tous les* $k = 1, \ldots, d$ :

- si les probabilités proposées $p_1, \ldots, p_d$ sont les bonnes, alors $Z$ suit une loi $\chi^2(d-1)$,
- si les probabilités proposées $p_1, \ldots, p_d$ ne sont pas les bonnes, alors $Z$ va être "beaucoup plus grand".

On a donc tout ce qu'il faut pour construire un test de :

$H_0 : \mathbb{P}(X = x_k) = p_k$ pour tout $k = 1 \ldots, d$,

(autrement dit : la loi de $X$ est bien donnée par $p_1, \ldots, p_k$)

contre

$H_1 : H_0$ n'est pas vraie

(autrement dit : la loi de $X$ n'est pas donnée par $p_1, \ldots, p_k$)

# Test du chi-deux d'adéquation : exemple

Nous allons traiter l'exemple numéro 3 :

On examine 1000 Français originaires du Pays basque, et on constate que 740 ont pour groupe sanguin $O$, 250 un groupe sanguin $A$, 5 un groupe sanguin $B$ et 5 un groupe sanguin $AB$. Sachant que la répartition dans l'ensemble de la population française est 43%, 45%, 9%, 3%, doit-on en conclure que la population basque présente une particularité dans la population française ?

# Test du chi-deux d'adéquation : exemple

On note $n = 1000$ et

$$p_O = 0{,}43,\ p_A = 0{,}45,\ p_B = 0{,}09,\ p_{AB} = 0{,}03$$

les probabilités "proposées".

On a

$$np_O = 430,\ np_A = 450,\ np_B = 90,\ np_{AB} = 30$$

qui sont tous $\geq 5$.

Par conséquent, si l'on pose

$$Z = \frac{(N_O - np_O)^2}{np_O} + \frac{(N_A - np_A)^2}{np_A} + \frac{(N_B - np_B)^2}{np_B} + \frac{(N_{AB} - np_{AB})^2}{np_{AB}}$$

alors si les probabilités proposées sont les bonnes, on a $Z \sim \chi^2(3)$ et autrement $Z$ doit être "très grand".

# Test du chi-deux d'adéquation : exemple détaillé

On rédige en sept points le test au niveau $\alpha = 1\%$.

1. **modèle :** les observations $X_1, \ldots, X_{1000}$ forment un 1000-échantillon d'une variable aléatoire qui peut prendre les valeurs O, A, B, AB.
2. **hypothèses :** on teste $H_0$ : $\mathbb{P}(X = O) = 0{,}43$, $\mathbb{P}(X = A) = 0{,}45$, $\mathbb{P}(X = B) = 0{,}09$, $\mathbb{P}(X = AB) = 0{,}03$ contre $H_1$ : ces égalités ne sont pas toutes vraies. On note $p_O = 0{,}43$, $p_A = 0{,}45$, $p_B = 0{,}09$, $p_{AB} = 0{,}03$.
3. **variable de test :** on utilise la variable

$$Z = \frac{(N_O - n\,p_O)^2}{n\,p_O} + \frac{(N_A - n\,p_A)^2}{n\,p_A} + \frac{(N_B - n\,p_B)^2}{n\,p_B} + \frac{(N_{AB} - n\,p_{AB})^2}{n\,p_{AB}}$$

4. **forme de la zone de rejet :** comme $n\,p_O = 430$, $n\,p_A = 450$, $n\,p_B = 90$, $n\,p_{AB} = 30$, toutes ces valeurs sont supérieures à 5, donc sous $H_0$, la variable $Z$ suit une loi $\chi^2(3)$ et sous $H_1$, la variable $Z$ prend des valeurs "très grandes" donc plus grandes que sous $H_0$. On va donc rejeter $H_0$ si $Z$ est "trop grand", on choisit une zone de rejet de la forme $[Z \geq z_0]$.

# Test du chi-deux d'adéquation : exemple détaillé

5. **seuil de la zone de rejet :** le seuil $z_0$ est le plus petit $z$ vérifiant $\mathbb{P}_{\mathsf{H}_0}(Z \geq z_0) \leq \alpha$. Comme $\mathbb{P}_{\mathsf{H}_0}(Z \geq z_0) = 1 - \mathbb{P}_{\mathsf{H}_0}(Z \leq z_0)$, on cherche le plus petit $z$ tel que $\mathbb{P}_{\mathsf{H}_0}(Z \leq z_0) \geq 1 - \alpha$. La table de valeurs numériques (ligne "$n = 3$", colonne "$p = 0{,}99$") donne $z_0 = 11{,}3449$.
6. **conclusion :** on calcule la valeur observée de $Z$

$$Z_{\text{obs}} = \frac{(740 - 1000 \times 0{,}43)^2}{1000 \times 0{,}43} + \frac{(250 - 1000 \times 0{,}45)^2}{1000 \times 0{,}45} + \frac{(5 - 1000 \times 0{,}09)^2}{1000 \times 0{,}09} + \frac{(5 - 1000 \times 0{,}03)^2}{1000 \times 0{,}03}$$

   donc $Z_{\text{obs}} \simeq 413{,}49$ et on rejette l'hypothèse $\mathsf{H}_0$ au niveau 1%.
7. ***p*-valeur :** la zone de rejet est de la forme $[Z \geq z_0]$ donc la $p$-valeur est $\mathbb{P}_{\mathsf{H}_0}(Z \geq Z_{\text{obs}})$. Tout ce que l'on peut dire avec la table est que $\mathbb{P}_{\mathsf{H}_0}(Z \geq Z_{\text{obs}}) \leq \mathbb{P}_{\mathsf{H}_0}(Z \geq 16{,}2662) = 0{,}1\%$ donc la $p$-valeur est inférieure à 0,1%.

# Test du chi-deux d'adéquation : restrictions

Le test du chi-deux d'adéquation que nous avons décrit a un certain nombre de limites, et en particulier :

1. on a besoin d'avoir $n \times p_k \geq 5$ pour tout $k = 1, \ldots, d$,
2. on a besoin que les variables $X$ soient discrètes, et qu'elles prennent seulement un nombre fini de valeurs,
3. on a besoin que la loi proposée pour l'hypothèse $H_0$ soit explicite.

On peut s'arranger pour appliquer ce test quand ces hypothèses ne sont pas vérifiées, et c'est ce que l'on va voir maintenant.

# Test du chi-deux d'adéquation : aménagements

1. Si l'on n'a pas $n \times p_k \geq 5$ pour tout $k = 1, \ldots, d$, alors on va regrouper des valeurs.
2. Si les variables $X$ sont discrètes mais prennent une infinité de valeurs, on va aussi les regrouper pour avoir un nombre fini de classes.

2'. Si les variables $X$ sont continues, alors on va regrouper les valeurs, par exemple par "fourchettes de valeurs".

   La manière dont on fait les regroupements dépend du contexte : à vous de faire des regroupements pertinents.

3. Si la loi proposée pour l'hypothèse $H_0$ n'est pas explicite mais qu'il manque un ou plusieurs paramètres, alors on remplace ce paramètre par une valeur estimée[5], et pour chaque paramètre estimé on fait baisser le nombre de degrés de liberté du $\chi^2$ : par exemple, $Z$ suivra une $\chi^2(d-2)$ sous $H_0$ si on a estimé un paramètre.

5. au sens où l'on utilise un estimateur comme décrit en page 162

# Exemple d'aménagement 1. et 2.

On s'intéresse au nombre d'accidents par jour sur la N118. On observe pour cela le nombre d'accidents quotidien sur 100 jours choisis au hasard et on veut tester l'hypothèse $H_0$ : $X$ suit une loi de Poisson de paramètre 2. On observe (en notant $N_0$ le nombre de jours avec 0 accidents, *etc.*)

$$N_0 = 21, N_1 = 30, N_2 = 29, N_3 = 14, N_4 = 4, N_5 = 1, N_6 = 1$$

(et $N_k = 0$ pour $k \geq 7$).

Problème : une loi de Poisson peut prendre une infinité de valeurs. On peut alors imaginer regrouper les catégories de valeurs en

| 0 | 1 | 2 | 3 | 4 | 5 | $\geq 6$ |
|---|---|---|---|---|---|---|

# Exemple d'aménagement 1. et 2.

Les probabilités correspondantes (les $p_k$ du théorème) sont alors :

- la proba qu'une v.a. de loi $\mathcal{P}(2)$ vaille 0,
- la proba qu'une v.a. de loi $\mathcal{P}(2)$ vaille 1,
- ...
- la proba qu'une v.a. de loi $\mathcal{P}(2)$ vaille 5,
- la proba qu'une v.a. de loi $\mathcal{P}(2)$ prenne des valeurs $\geq 6$,

que l'on peut déterminer grâces aux tables de la loi de Poisson $\mathcal{P}(2)$.

On a donc le tableau des "$p_k$", des "$n\,p_k$" et des "$N_k$" :

| $k$ | 0 | 1 | 2 | 3 | 4 | 5 | $\geq 6$ |
|---|---|---|---|---|---|---|---|
| $p_k$ | 0,1353 | 0,2707 | 0,2707 | 0,1804 | 0,0902 | 0,0361 | 0,0166 |
| $np_k$ | 13,53 | 27,07 | 27,07 | 18,04 | 9,02 | 3,61 | 1,66 |
| $N_k$ | 21 | 30 | 29 | 14 | 4 | 1 | 1 |

# Exemple d'aménagement 1. et 2.

mais dans ce cas les "$np_k$" ne sont pas tous $\geq 5$. On regroupe donc les valeurs 5 et $\geq 6$ en une classe $\geq 5$, d'où le nouveau tableau

| $k$ | 0 | 1 | 2 | 3 | 4 | $\geq 5$ |
|---|---|---|---|---|---|---|
| $p_k$ | 0,1353 | 0,2707 | 0,2707 | 0,1804 | 0,0902 | 0,0527 |
| $np_k$ | 13,53 | 27,07 | 27,07 | 18,04 | 9,02 | 5,27 |
| $N_k$ | 21 | 30 | 29 | 14 | 4 | 2 |

dans ce cas tous les "$np_k$" sont $\geq 5$ et on va pouvoir faire le test.

La loi du $Z$ sous l'hypothèse $H_0$ sera une $\chi^2(5)$ car le $d$ considéré est le nombre de "cases" et les cases sont ici au nombre de six : 0, 1, 2, 3, 4 et $\geq 5$.

# Exemple d'aménagement 2'.

On s'intéresse à la taille des hommes jeunes français et on veut tester si la taille d'un tel Français choisi au hasard suit bien une loi normale $\mathcal{N}(179,7^2)$.

On peut par exemple regrouper les tailles des individus observés en "moins de 165", "entre 165 et 170",...,"entre 185 et 190", "plus de 190".

Les probabilités correspondantes (les $p_k$ du théorème) sont alors :

- la proba qu'une v.a. de loi $\mathcal{N}(179,7^2)$ prenne des valeurs $< 165$,
- la proba qu'une v.a. de loi $\mathcal{N}(179,7^2)$ prenne des valeurs dans $[165,170[$,
- ...
- la proba qu'une v.a. de loi $\mathcal{N}(179,7^2)$ prenne des valeurs dans $[185,190[$,
- la proba qu'une v.a. de loi $\mathcal{N}(179,7^2)$ prenne des valeurs $\geq 190$,

# Exemple d'aménagement 2'.

On obtient après calculs

(de $\mathbb{P}(X \leq 165)$, $\mathbb{P}(165 \leq X \leq 170) \ldots$ pour $X \sim \mathcal{N}(179, 7^2)$)

la table

| $k$ | $< 165$ | [165,170[ | [170,175[ | [175,180[ | [180,185[ | [185,190[ | $\geq 190$ |
|---|---|---|---|---|---|---|---|
| $p_k$ | 0,0227 | 0,0765 | 0,1846 | 0,2729 | 0,2475 | 0,1376 | 0,0580 |

encore une fois il faudra faire attention à avoir tous les "$np_k$" supérieurs à 5 !

par exemple

- si $n = 250$ c'est déjà le cas, on a alors $d = 7$,
- si $n = 100$ il faudrait regrouper "moins de 165" et "entre 165 et 170" en une classe "moins de 170" et on aurait alors $d = 6$.

.

# Exemple d'aménagement 3.

On reprend l'exemple de la N118 mais cette fois-ci on veut tester l'hypothèse $H_0$ : $X$ suit une loi de Poisson contre $H_1$ : $X$ ne suit pas une loi de Poisson.

Comme on n'a pas le paramètre de la loi de Poisson, on va *l'estimer*. Comme le paramètre d'une loi de Poisson est égal à sa moyenne, on va utiliser la moyenne empirique (le $\overline{X}_n$) comme estimateur, et on trouve

$$\frac{1}{100}(21 \times 0 + 30 \times 1 + 29 \times 2 + 14 \times 3 + 4 \times 4 + 1 \times 5 + 1 \times 6) = 1{,}57.$$

On fait donc le test avec une loi de Poisson $\mathcal{P}(1{,}57)$ et on enlèvera un degré de liberté à la loi du chi-deux quand on cherchera le seuil de la zone de rejet (si par exemple on regroupe les valeurs en six "cases" comme précédemment alors sous l'hypothèse $H_0$, la loi de $Z$ sera une $\chi^2(4)$ et non $\chi^2(5)$).

# Test du chi-deux d'indépendance : introduction

Supposons que l'on a $(X_1, Y_1), \ldots, (X_n, Y_n)$ un $n$-échantillon d'observations indépendantes, où les $X$ peuvent prendre les valeurs $a_1, \ldots, a_d$ et les $Y$ les valeurs $b_1, \ldots, b_e$ respectivement. On interprète $(X_k, Y_k)$ comme l'observation des caractéristiques $X$ et $Y$ chez le $k$-ième individu.

**Exemple :** pour le $k$-ième pommier, on note $X_k$ le traitement qu'il a reçu (qui vaut donc $A$, $B$, $C$ ou $D$) et $Y_k$ le fait qu'il ait été productif ou non (qui vaut donc $P$ ou $NP$).

(on a donc $n = 1505$, $d = 4$, $e = 2$)

On veut tester si les caractéristiques $X$ et $Y$ sont indépendantes.

# Test du chi-deux d'indépendance : notations

On note

- $N_k$ le nombre de fois où la valeur $a_k$ est observée parmi $X_1, \ldots, X_n$,
- $N^\ell$ le nombre de fois où la valeur $b_\ell$ est observée parmi $Y_1, \ldots, X_n$,
- $N_k^\ell$ le nombre de fois où la paire de valeurs $(a_k, b_\ell)$ est observée parmi $(X_1, Y_1), \ldots, (X_n, Y_n)$.

On a forcément

$$N_k = \sum_{\ell=1}^{e} N_k^\ell, \qquad N^\ell = \sum_{k=1}^{d} N_k^\ell, \qquad \sum_{k=1}^{d} N_k = \sum_{\ell=1}^{e} N^\ell = \sum_{k=1}^{d} \sum_{\ell=1}^{e} N_k^\ell = n.$$

# Test du chi-deux d'indépendance : notations

**Exemple :** On notera

- $N_P$, $N_{NP}$ le nombre de pommiers productifs et non productifs
- $N^A$, $N^B$, $N^C$, $N^D$ le nombre qui ont reçu le traitement $A$, $B$, $C$, $D$ respectivement,
- $N_P^A$ le nombre de pommiers qui ont reçu le traitement $A$ et sont productifs, *etc.*

et le tableau

| | $A$ | $B$ | $C$ | $D$ |
|---|---|---|---|---|
| $P$ | 156 | 113 | 128 | 185 |
| $NP$ | 203 | 266 | 258 | 196 |

nous donne

| | $A$ | $B$ | $C$ | $D$ | total |
|---|---|---|---|---|---|
| $P$ | $N_P^A = 156$ | $N_P^B = 113$ | $N_P^C = 128$ | $N_P^D = 185$ | $N_P = 582$ |
| $NP$ | $N_{NP}^A = 203$ | $N_{NP}^B = 266$ | $N_{NP}^C = 258$ | $N_{NP}^D = 196$ | $N_{NP} = 923$ |
| total | $N^A = 359$ | $N^B = 379$ | $N^C = 386$ | $N^D = 381$ | $n = 1505$ |

# Test du chi-deux d'indépendance : théorème

Le résultat théorique qui fait fonctionner le test d'indépendance est le suivant :

**Théorème**

si l'on note

$$Z = \sum_{k=1}^{d} \sum_{\ell=1}^{e} \frac{(N_k^\ell - \frac{N_k N^\ell}{n})^2}{\frac{N_k N^\ell}{n}}$$

alors

- si les caractères $X$ et $Y$ sont indépendants, alors pour $n$ grand, $Z$ suit (à peu près) une loi du chi-deux à $(d-1)\times(e-1)$ degrés de liberté,
- si ces caractères ne sont pas indépendants, alors $Z \to \infty$ quand $n \to \infty$, donc pour $n$ grand $Z$ est "très grand".

Le **critère pratique** pour "$n$ grand" est que $\frac{N_k N^\ell}{n} \geq 5$ pour tous les $k$ et $\ell$.

# Test du chi-deux d'indépendance : théorème

Autrement dit, si $\frac{N_k N^\ell}{n} \geq 5$ pour tous les $k$ et $\ell$ :

- si les caractères $X$ et $Y$ sont indépendants, alors $Z$ suit une loi $\chi^2\big((d-1)\times(e-1)\big)$,
- si les caractères $X$ et $Y$ ne sont pas indépendants, alors $Z$ va être "beaucoup plus grand".

On a donc tout ce qu'il faut pour construire un test de :

$\mathrm{H}_0$ : $X$, $Y$ sont indépendants

contre

$\mathrm{H}_1$ : $X$, $Y$ ne sont pas indépendants.

# Test du chi-deux d'indépendance : exemple

Nous allons traiter l'exemple des pommiers :

On observe 1505 pommiers qui ont tous reçu un traitement parmi quatre possibles (notés $A$, $B$, $C$ et $D$) et on note s'ils ont été productifs (noté $P$) ou non productifs (noté $NP$). Pour le pommier $k$, on note $X_k$ sa productivité et $Y_k$ le traitement qu'il a reçu. On a donc $n = 1505$ et les valeurs possibles des $X$ sont $P$ et $NP$, les valeurs possibles des $Y$ sont $A$, $B$, $C$ et $D$. Les résultats sont résumés dans le tableau suivant :

| | $A$ | $B$ | $C$ | $D$ |
|---|---|---|---|---|
| $P$ | 156 | 113 | 128 | 185 |
| $NP$ | 203 | 266 | 258 | 196 |

# Test du chi-deux d'indépendance : exemple

Le tableau ci-dessus donnes les valeurs de $N_P^A$, $N_P^B$, etc.

On en déduit les valeurs de $N_P$, $N_{NP}$, $N^A$, $N^B$, $N^C$, $N^D$ :

| | $A$ | $B$ | $C$ | $D$ | total |
|---|---|---|---|---|---|
| $P$ | 156 | 113 | 128 | 185 | 582 |
| $NP$ | 203 | 266 | 258 | 196 | 923 |
| total | 359 | 379 | 386 | 381 | 1505 |

puis le tableau des $N_k N^\ell / n$ :

| | $A$ | $B$ | $C$ | $D$ |
|---|---|---|---|---|
| $P$ | 138,83 | 146,56 | 149,27 | 147,34 |
| $NP$ | 220,17 | 232,44 | 236,73 | 233,66 |

(par exemple, 138,83 c'est $359 \times 582/1505$).

# Test du chi-deux d'indépendance : exemple

Les $N_k N^\ell / n$ sont tous $\geq 5$. Par conséquent si l'on pose

$$Z = \sum_{k=P,NP} \sum_{\ell=A,B,C,D} \frac{(N_k^\ell - \frac{N_k N^\ell}{n})^2}{\frac{N_k N^\ell}{n}}$$

alors si $X$ et $Y$ sont indépendants, on a $Z \sim \chi^2(3)$ (où 3 est $(2-1)\times(4-1)$) et autrement $Z$ doit être "très grand".

# Test du chi-deux d'indépendance : exemple détaillé

On rédige en sept points le test au niveau $\alpha = 2\%$.

1. **modèle :** les observations $(X_1, Y_1) \ldots, (X_{1505}, Y_{1505})$ forment un 1505-échantillon d'un couple de variables aléatoires $(X, Y)$, où $X$ peut prendre les valeurs $P$ (productif) et $NP$ (non productif) et $Y$ les valeurs $A$, $B$, $C$, $D$ (le type de traitement suivi).
2. **hypothèses :** on teste $H_0$ : $X$ et $Y$ sont indépendants contre $H_1$ : $X$ et $Y$ ne sont pas indépendants.
3. **variable de test :** on utilise la variable

$$Z = \sum_{k=P,NP} \sum_{\ell=A,B,C,D} \frac{(N_k^\ell - \frac{N_k N^\ell}{n})^2}{\frac{N_k N^\ell}{n}}$$

4. **forme de la zone de rejet :** toutes les valeurs des $\frac{N_k N^\ell}{n}$ sont supérieures à 5, donc sous $H_0$, la variable $Z$ suit une loi $\chi^2\big((2-1)\times(4-1)\big)$ et sous $H_1$, la variable $Z$ prend des valeurs "très grandes" donc plus grandes que sous $H_0$. On va donc rejeter $H_0$ si $Z$ est "trop grand" et on choisit donc une zone de rejet de la forme $[Z \geq z_0]$.

# Test du chi-deux d'indépendance : exemple détaillé

5. **seuil de la zone de rejet :** le seuil $z_0$ est le plus petit $z$ vérifiant $\mathbb{P}_{\mathsf{H}_0}(Z \geq z) \leq \alpha$. Comme $\mathbb{P}_{\mathsf{H}_0}(Z \geq z) = 1 - \mathbb{P}_{\mathsf{H}_0}(Z \leq z)$, on cherche le plus petit $z$ tel que $\mathbb{P}_{\mathsf{H}_0}(Z \leq z) \geq 1 - \alpha$. La table de valeurs numériques (ligne "$n = 3$", colonne "$p = 0{,}98$") donne $z_0 = 9{,}8374$.
6. **conclusion :** on calcule la valeur observée de $Z$

$$Z_{\text{obs}} = \frac{(156-138{,}83)^2}{138{,}83} + \frac{(203-220{,}17)^2}{220{,}17} + \frac{(113-146{,}56)^2}{146{,}56} + \frac{(266-232{,}44)^2}{232{,}44} + \frac{(128-149{,}27)^2}{149{,}27} + \frac{(258-236{,}73)^2}{236{,}73} + \frac{(185-149{,}27)^2}{149{,}27} + \frac{(196-233{,}66)^2}{233{,}66}$$

   donc $Z_{\text{obs}} \simeq 36{,}63$ et on rejette l'hypothèse $\mathsf{H}_0$ au niveau 2%.
7. ***p*-valeur :** la zone de rejet est de la forme $[Z \geq z_0]$ donc la $p$-valeur est $\mathbb{P}_{\mathsf{H}_0}(Z \geq Z_{\text{obs}})$. Tout ce que l'on peut dire avec la table est que $\mathbb{P}_{\mathsf{H}_0}(Z \geq Z_{\text{obs}}) \leq \mathbb{P}_{\mathsf{H}_0}(Z \geq 16{,}2662) = 0{,}1\%$ donc la $p$-valeur est inférieure à 0,1%.

# Test du chi-deux d'indépendance : restrictions

Le test du chi-deux d'indépendance souffre des mêmes limitations que le test du chi-deux d'adéquation :

1. on a besoin que les $N_k N^\ell / n$ soient tous $\geq 5$,
2. on a besoin que les variables $X$ et $Y$ soient discrètes, et qu'elles prennent seulement un nombre fini de valeurs.

mais on peut faire les mêmes aménagements que pour le test du chi-deux d'adéquation :

1. Si l'on n'a pas $N_k N^\ell / n \geq 5$ pour tous $k$ et $\ell$, on regroupe des valeurs.
2. Si les variables $X$, $Y$ sont discrètes mais prennent une infinité de valeurs, on les regroupe pour avoir un nombre fini de classes.

2'. Si les variables $X$, $Y$ sont continues, alors on va regrouper les valeurs, par exemple par "fourchettes de valeurs".

La manière dont on fait les regroupements dépend du contexte : à vous de faire des regroupements pertinents.

# Test du chi-deux d'homogénéité

Il existe un troisième type de test proche des deux précédents.

Il permet de traiter par exemple :

On examine 500 Français originaires des Bouches-du-Rhône et 750 Français originaires de Corse. On obtient les répartitions suivantes :

| | $O$ | $A$ | $B$ | $AB$ |
|---|---|---|---|---|
| BdR | 223 | 218 | 34 | 25 |
| Corse | 382 | 317 | 23 | 28 |

Doit-on en conclure que les populations des Bouches-du-Rhône et de Corse ont les mêmes répartions de groupes sanguins ?

sans se demander si ces répartitions sont les mêmes que celle dans l'ensemble des Français – donc sans avoir besoin de connaître cette dernière.

# Test du chi-deux d'homogénéité

Ce test d'homogénéité revient à tester l'indépendance de la caractéristique "groupe sanguin" et de la caractéristique "origine".

Le test est donc exactement un test d'indépendance basé sur le tableau

| | $O$ | $A$ | $B$ | $AB$ |
|---|---|---|---|---|
| BdR | 223 | 218 | 34 | 25 |
| Corse | 382 | 317 | 23 | 28 |