HAI718 Probabilités et statistiques

TD2 : Tests statistiques sur un paramètre d'intérêt – Correction

1 Introduction et problématique

Un industriel veut lancer sur le marché un nouveau produit issu de l'agriculture biologique que l'on note *Produit A*. Ce produit est bien entendu en concurrence avec d'autres produits.

Après une étude financière, les services comptables indiquent à cet industriel que pour le lancement d'un produit soit rentable, il faut qu'il soit vendu à plus de 300000 exemplaires par mois. La population ciblée par l'industriel est une population de taille $N = 2$ millions. Autrement dit, pour que le produit soit mis sur le marché, il faut que la proportion d'acheteurs potentiels soit supérieure à 15 %.

Exercice 1 *Dégagez le* **paramètre d'intérêt** *de l'étude et proposez-en une notation.*

Le paramètre d'interêt est le taux de gens intéressés, notons-le T. Il est égal au nombre de gens intéressés divisé par la population totale.

1.1 La solution idéale

Exercice 2 *Quelle serait la solution idéale pour répondre à cette problématique ? Pourquoi cette solution paraît-elle peu envisageable ? Quelle serait une solution réalisable ?*

La solution idéale consisterait à interroger toute la population. C'est trop cher et trop long. Une solution réalisable : un sondage d'une plus petite partie de la population.

1.2 Une solution réalisable

La solution réalisable consiste à interroger un nombre moins important de consommateurs potentiels. Supposons que l'on puisse appliquer la procédure précédente sur une sous-population de taille $n = 100$ consommateurs potentiels.

Exercice 3 *Donnez une* **estimation** *du paramètre d'intérêt et une notation pour cette estimation ? En général, la valeur de cette estimation et la "vraie" valeur du paramètre d'intérêt sont différentes. Pourriez-vous en donnez les explications ?*

Comment conluriez-vous quant au problème posé par l'industriel ?

Une estimation du taux de gens intéressés est le taux de gens intéressés parmi les gens interrogés. On le note $\hat{T}$. Les raisons qui font différer T et $\hat{T}$ sont de plusieurs ordres :

— échantillon trop petit

— échantillon déséquilibré et donc non représentatif du comportement global (exemple si on n'interroge que des hommes, ou que des femmes, ou mauvaise répartition selon l'âge, etc.).
— facteur extérieur contingent aux conditions du sondage (conditions climatiques, climat social tendu, etc.).
— la formulation de la question du sondage (plus ou moins floue sur le fait d'acheter. Par exemple si on pose la question "Seriez-vous intéressé de découvrir notre lessive ?" ou "Achéterez-vous cette lessive ?", la réponse ne va pas être la même...)

1.2.1 Formalisation mathématique du problème

• La population totale

Dans cette étude, la population totale correspond à la population des consommateurs potentiels. On la note $\Omega = \{\omega_1, \ldots, \omega_N\}$ où ω_i désigne le $i$ème individu de la population totale Ω.

Soit X la fonction qui à tout individu ω_i de la population totale associe la réponse de l'individu ω_i à la question : "achèteriez-vous le *Produit A* ?". En fait, on utilise le codage suivant : $X(\omega_i) = 1$ pour une réponse par **oui** et à $X(\omega_i) = 0$ pour une réponse par non.

Exercice 4 *Suivant ces notations, que vaut alors le paramètre d'intérêt ?*

$$T = \frac{\sum_{i=1}^{N} X(\omega_i)}{N}.$$

• L'échantillon

La fonction X est associée à une *variable aléatoire* que l'on notera également X.

Exercice 5 *Que vaut $P(X = 1)$? Quel est le nom de cette variable aléatoire et l'expression de sa densité ?*

$P(X = 1) = T$. C'est une variable de Bernoulli de paramètre T, et on a : $P(X = 0) = 1 - T$.

On désigne par $\mathbf{e} = \{e_1, \ldots, e_n\}$ l'échantillon de taille n. À partir de celui-ci, on obtient alors le vecteur d'observations $\mathbf{x}^A = (x_1, \ldots, x_n)$ pour les variables $\mathbf{X}^A = (X_1, \ldots, X_n)$. Cela signifie simplement que $X(e_i) = x_i$ pour $i = 1, \ldots, n$.

Exercice 6 *Quel est l'estimateur du paramètre d'intérêt ? Quelle est son estimation ?*

$$\hat{T} = \frac{\sum_{i=1}^{n} X_i}{n}$$

qui suit une loi binomiale. Son estimation est

$$\hat{T}(e) = \frac{\sum_{i=1}^{n} x_i}{n}.$$

La notation $\hat{T}(e)$ indique clairement que cette quantité, contrairement à $\hat{T}$, est uniquement dépendante des informations relatives aux individus de la sous-population $\mathbf{e}$.

1.3 Décision devant être prise

Exercice 7 *À l'aide des informations et des notations précédentes, proposez une règle de décision qui permettrait de donner une réponse à l'industriel. Argumentez et critiquez votre proposition. Qu'en pensez-vous ?*

Une règle de décision simple serait de calculer $\hat{T}(e)$ sur un échantillon donné de population, et si cette valeur est $< 15\%$, on ne lance pas le produit, sinon on le lance. Mais dans cette décision, on est suceptible de faire des erreurs car l'estimation peut être différente de la réalité. Ce qu'il faut réussir à quantifier, c'est :

— de combien on risque de se tromper (et dans quel sens)
— pour un risque donné, quel est le seuil de décision que l'on prend (à la place de 15% éventuellement)

1.4 Mise en évidence d'un phénomène aléatoire

Voici les résultats de trois études menées pour le lancement du produit dans trois pays différents. Dans chaque cas, l'étude est réalisée sur un échantillon de taille $n = 1000$ individus. Voici ce que l'on observe :

- **Pays 1** : le nombre de consommateurs potentiels est 123
- **Pays 2** : le nombre de consommateurs potentiels est 155
- **Pays 3** : le nombre de consommateurs potentiels est 176

Exercice 8 *La question à laquelle vous devez répondre est : Dans quel(s) pays l'industriel doit-il lancer le produit ?*

Le taux est 12.3% dans le pays 1, 15.5% dans le pays 2 et 17.6% dans le pays 3. Si on prend la règle de décision précédente, on lancerait le produit dans les pays 2 et 3.

1.4.1 Décision prise par le statisticien

Le statisticien conseillerait de ne lancer son produit que pour le troisième pays, et de ne pas le lancer pour les pays 1 et 2.

Exercice 9 *Qu'en pensez-vous ? Quels pays auriez-vous choisi ?*

Le statisticien a certainement quantifié le seuil à prendre en fonction du risque choisi. Autrement dit, il a calculé que la possibilité de se tromper lorsque le taux estimé est proche de 15% comme l'est 15.5% est trop grand pour accepter l'hypothèse que la population intéressé a un taux $\geq 15\%$. Nous allons nous mettre dans sa peau par la suite...

Commentaires supplémentaires La règle du statisticien diffère de celle que l'on prendrait naïvement. Il faut se rappeler que toute décision basée sur un échantillon est entachée d'erreur. La question se pose alors de savoir s'il est possible d'évaluer cette erreur.

On est amené à penser que l'inévitable erreur de décision est de nature aléatoire. Plus précisément, il est quasi certains que le procédé d'extraction de l'échantillon est un phénomène aléatoire et que n'importe quelle

règle de décision conduira à des erreurs de décision. Le but de la statistique inférentielle est entre autre chose de proposer un outil d'aide à la décision , appelé test d'hypothèses, permettant de contrôler les erreurs de décisions. Le développement d'un tel outil repose sur la connaissance et les propriétés du phénomène aléatoire relatif à l'extraction de l'échantillon. Ce procédé est appelé échantillonnage.

Il existe différentes stratégies d'échantillonnage. Les deux plus connues sont l'échantillonnage sans et avec remise dans la population totale. En théorie, cette distinction implique des règles de décision différentes. Cependant, lorsque la taille de la population N est très grande devant la taille de l'échantillon n ($N \geq 100 * n$), les résultats théoriques sont très proches. Nous ne tiendrons pas compte de cette distinction et supposerons que tout les tirages sont effectués avec remise.

1.4.2 Les types d'erreur

Un agronome se procure des grains chez un fournisseur qui produit des sacs de deux qualités différentes : des sacs de bonne qualité qui contiennent 6% de grains "défectueux" (qui ne germent pas), et des sacs de mauvaise qualité contenant 10% de grains défectueux. On veut déterminer la qualité du sac reçu. On pose le test d'hypothèses suivant :

$$H_0 : p = 10\% \text{ contre } H_1 : p = 6\%.$$

L'effectif de l'échantillon pour réaliser le test est $n = 100$ grains. On note X_i une variable aléatoire de Bernoulli de paramètre p (notée $\mathcal{B}e(p)$) telle que :
$X_i = 1$ si le grain est défectueux et $P(X_i = 1) = p$
$X_i = 0$ si le grain n'est pas défectueux et $P(X_i = 0) = 1 - p$.

La loi de $X = \sum_{i=1}^{n} X_i$ est une loi Binomiale de paramètres n et p (notée $\mathcal{B}(n, p)$) avec :

$$P(X = k) = C_n^k p^k (1-p)^{(n-k)} \text{ et } C_n^k = \frac{n!}{k!(n-k)!}.$$

Pour $n > 30$, on peut approximer la loi de X/n par une loi normale et

$$X/n \sim \mathcal{N}\left(\mu = p, \sigma^2 = \frac{p(1-p)}{n}\right).$$

1.4.3 Les erreurs de première et seconde espèce

Avec un seuil de décision $s = 8\%$, on définit la règle de décision suivante :
si $\hat{p} \leq s$, on accepte H_1 et on abandonne H_0
si $\hat{p} > s$, on conserve H_0

Exercice 10 *Suivant cette règle de décision et en utilisant l'approximation de la loi X/n, quels sont les risques*

1. *de première espèce : $\alpha = P(\text{accepter } H_1 \text{ à tort})$*
2. *de deuxième espèce : $\beta = P(\text{conserver } H_0 \text{ à tort})$.*

Afin de répondre, on représentera graphiquement la distribution du test sous les deux hypothèses H_0 et H_1.

On s'aide de R, comme on l'a déjà fait dans le TP2, ou bien des tables de la loi normale, après avoir centré et réduit la variable aléatoire. Le risque de première espèce est $P(X_{H_0} < s)$.

```
> pnorm(0.08,0.1,sqrt((0.1*0.9)/100))
[1] 0.2524925
```

C'est-à-dire un risque d'un peu plus de 25%

Le risque de deuxième espèce est $P(X_{H_1} \geq s)$:

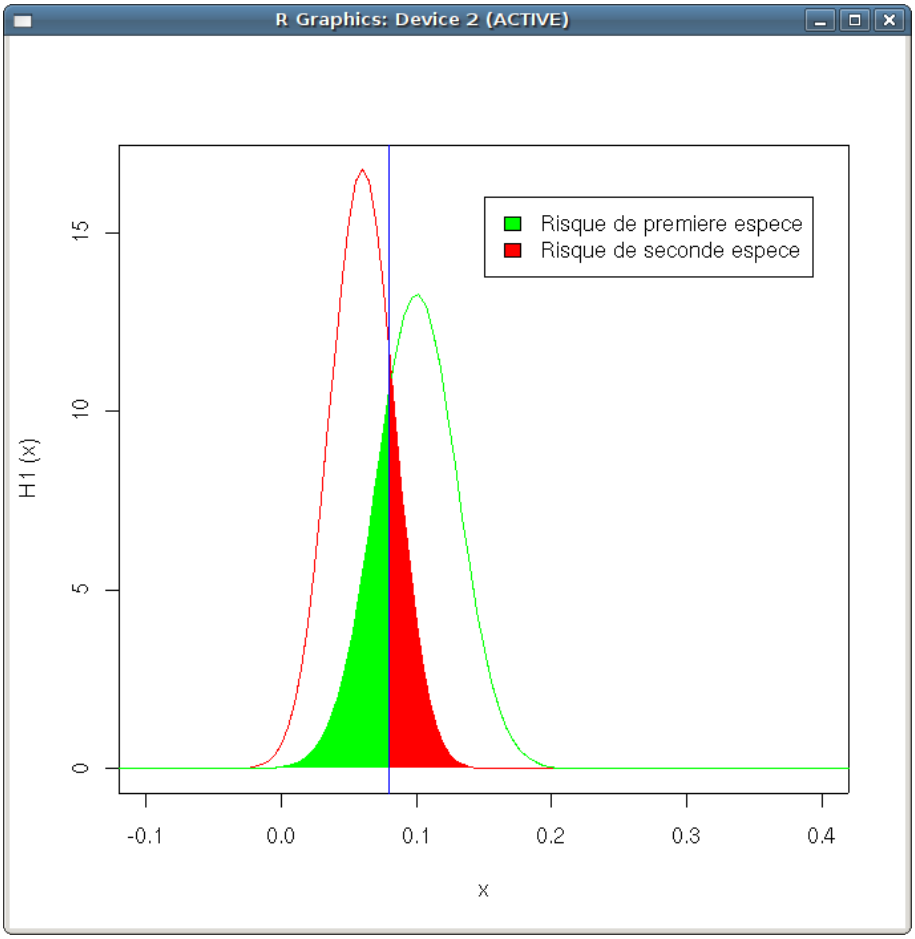
```
> pnorm(0.08,0.06,sqrt((0.06*0.94)/100),lower.tail=F)
[1] 0.1998515
```

soit un peu moins de 20%.

La représentation graphique correspondant à ces deux lois :

```
> H0 <- function (x) dnorm(x,0.1,sqrt(0.1*0.9/100))
> H1 <- function (x) dnorm(x,0.06,sqrt(0.06*0.94/100))
> plot(H1,-0.1,0.4,col="red")
> curve(H0,col="green",add=T)
> x <- seq(-0.1,0.4,length=501)
> polygon(c(-0.1,x[1:181],0.08),c(0,H0(x[1:181]),0),border=0,col="green")
> polygon(c(0.08,x[181:501],0.4),c(0,H1(x[181:501]),0),border=0,col="red")
> legend(0.15,16,fill=c("green","red"),legend=c("Risque de premiere espece",
  "Risque de seconde espece"))
> abline(v=0.08,col="blue")
```

Voici le graphique produit :



Admettons que l'on pose les hypothèses suivantes :

$$H_0 : p = 6\% \text{ contre } H_1 : p = 10\%.$$

Exercice 11 *Le résultat obtenu à la question précédente serait-il différent ? Expliquer pourquoi, éventuellement à l'aide de graphiques.*

Les rôles de α et β sont échangés.

1.5 Choix du risque α

Exercice 12 *On fixe $\alpha = 0.05$. Déterminer la valeur de s correspondante. Calculer la valeur du risque β. Qu'en concluez-vous ?*

La valeur trouvée pour β est proche de 0.65. La nouvelle valeur de $\beta = 65,3\%$ n'est évidemment pas acceptable... C'est certainement que l'échantillonnage est trop petit pour que l'on puisse discriminer la loi sous H_0 et la loi sous H_1.

1.6 Choix de α et β

Exercice 13 *On fixe $\alpha = 0.05$ et $\beta = 0.05$. Quelles sont les valeurs de s et de n correspondant à ces risques ?*

On a comme dans le premier exercice un système de deux équations à deux inconnues à résoudre. En effet, si on note $p_0 = 0.06$ et $p_1 = 0.1$, si X_0, X_1 et U sont des variables aléatoires de lois respectives $\mathcal{N}(p_0, \frac{p_0(1-p_0)}{n})$ (loi sous H_0), $\mathcal{N}(p_1, \frac{p_1(1-p_1)}{n})$ (loi sous H_1), et $\mathcal{N}(0,1)$, on a :

$$\alpha = P(X_0 < s) = P(U < \frac{s-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}) \text{ et } \beta = P(X_1 > s) = P(U > \frac{s-p_1}{\sqrt{\frac{p_1(1-p_1)}{n}}})$$

Les données du problème sont α et β, les inconnues sont s et n. Soient t_α tel que $\alpha = P(U < t_\alpha)$, et t_β tel que $\beta = P(U > t_\beta)$. Alors :

$$\left\{ \begin{array}{rcl} t_\alpha & = & \frac{s-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \\ t_\beta & = & \frac{s-p_1}{\sqrt{\frac{p_1(1-p_1)}{n}}} \end{array} \right.$$

Donc

$$\left\{ \begin{array}{rcl} t_\alpha \sqrt{\frac{p_0(1-p_0)}{n}} & = & s-p_0 \\ t_\beta \sqrt{\frac{p_1(1-p_1)}{n}} & = & s-p_1 \end{array} \right.$$

On traite ce système comme un système linéaire de deux équations à deux inconnues s et $\frac{1}{\sqrt{n}}$, ce qui nous donne :

$$\left\{ \begin{array}{rcl} \frac{1}{\sqrt{n}} & = & \frac{p_1-p_0}{t_\alpha \sqrt{p_0(1-p_0)} - t_\beta \sqrt{p_1(1-p_1)}} \\ s & = & p_0 + t_\alpha \sqrt{p_0(1-p_0)} \left(\frac{p_1-p_0}{t_\alpha \sqrt{p_0(1-p_0)} - t_\beta \sqrt{p_1(1-p_1)}} \right) \end{array} \right.$$

Soit finalement :

$$\left\{ \begin{array}{rcl} n & = & \left(\frac{t_\alpha \sqrt{p_0(1-p_0)} - t_\beta \sqrt{p_1(1-p_1)}}{p_1-p_0} \right)^2 \\ s & = & p_0 + t_\alpha \sqrt{p_0(1-p_0)} \left(\frac{p_1-p_0}{t_\alpha \sqrt{p_0(1-p_0)} - t_\beta \sqrt{p_1(1-p_1)}} \right) \end{array} \right.$$

On trouve pour n une valeur de 488.5. Comme n est un entier, on lui donnera la valeur 489. On recalcule s en fonction de ce nouveau n :

$$s = p_0 + \frac{1}{\sqrt{n}} t_\alpha \sqrt{p_0(1-p_0)}$$

Cela ne change pas fondamentalement la valeur du seuil, que nous fixerons à $7,77\%$.

2 Le problème des naissances

« En 1660, le négociant anglais Graunt découvrait, d'après les registres des baptêmes anglicans, un rapport de 105 garçons pour 100 filles, rapport présentant deux étranges particularités : sa remarquable constance d'un pays à l'autre et d'une époque à l'autre et son accroissement après guerre ou les famines ! Double mystère longtemps attribué à un « ordre divin » mais qu'aucun chercheur n'a pu à ce jour éclaircir »[1]

Voici ci-dessous des données recueillies par des étudiants du CUST (génie biologique) auprès de l'INSEE de Clermont-Ferrand.

	1969	1970	1971	1972
Allier : Garçons	2653	2735	2730	2716
Filles	2621	2521	2631	2640
Total	5274	5256	5361	5356
Cantal : Garçons	1290	1200	1257	1258
Filles	1268	1243	1246	1223
Total	2558	2443	2503	2481
Haute-Loire : Garçons	1504	1477	1469	1386
Filles	1437	1410	1443	1375
Total	2941	2887	2912	2761
Puy-de-Dôme : Garçons	4468	4567	4806	4864
Filles	4164	4263	4551	4664
Total	8632	8830	9357	9528
Auvergne : Garçons	9915	9979	10262	10224
Filles	9490	9437	9871	9902
Total	19405	19416	20133	20126

Au total 79080 naissances et parmi elles 40380 naissances de garçons.

Exercice 14 *Ces résultats confirment-ils ceux de Graunt ? À savoir :*

1. *Rejet de l'hypothèse qu'à la naissance il y a autant de filles que de garçons.*

   Hypothèses
   La variable aléatoire représentant le sexe d'un bébé qui naît suit une loi de Bernoulli de paramètre p. On a :

$$\begin{cases} P(\text{Garçon}) &= p \\ P(\text{Fille}) &= 1-p \end{cases}$$

1. D'après Schwartz (1993) Le jeu de la Science et du Hasard. Ed. Flammarion.

C'est sur la valeur de p que l'on fait des hypothèses. Dans le cas présent, l'hypothèse à tester est : H_0 : "$p = \frac{1}{2}$".
L'hypothèse alternative, au vu des résultats du tableau précédent, est la suivante : H_1 : "$p > \frac{1}{2}$". On réalise donc un **test unilatéral**.
Le taux de naissances de garçons dans une population de taille n, noté $Y = \frac{\sum_{i=1}^{n} X_i}{n}$, avec chaque X_i suivant une loi de Bernoulli de paramètre p (X_i vaut 1 si c'est un garçon, 0 sinon), suit, à un facteur $\frac{1}{n}$ près, une loi binomiale à n expériences et de paramètre p. Comme n est grand, on peut approximer cette loi binomiale par une loi normale de moyenne p et d'écart-type $\sqrt{\frac{p(1-p)}{n}}$.
Sous l'hypothèse H_0, cela signifie donc que le taux de naissances de garçons dans notre échantillon suit une loi $\mathcal{N}\left(\frac{1}{2}, \sqrt{\frac{\frac{1}{2}\left(1-\frac{1}{2}\right)}{n}}\right)$.

Zone critique
On veut que la probabilité que le taux de naissances de garçons soit supérieur à la valeur critique ne dépasse pas un certain risque, lorsque Y suit la loi sous l'hypothèse H_0. Par exemple, prenons un risque de 5%. On veut connaître le seuil s tel que $P(Y > s) = 5\%$.
Soit on demande directement la valeur à R pour les paramètres de la loi de Y, soit on normalise la variable aléatoire puis on lit dans les tables la valeur de $u = \frac{s-p}{\sqrt{\frac{p(1-p)}{n}}}$, puis on en déduit la valeur de s.

```
> qnorm(0.05,1/2,sqrt(1/(4*79080)),lower.tail=F)
[1] 0.5029246
```

Le seuil sera donc 0,5029246.

Décision Si la valeur prise par l'échantillon est inférieure au seuil, on accepte H_0, sinon on rejette H_0. Ici la valeur prise par l'échantillon est $\frac{40380}{79080} = 0.5106222$, c'est supérieur au seuil, donc on rejette H_0 (au risque de 5%).

Remarque : Si on avait pris un risque de 1%, le seuil aurait été :

```
> qnorm(0.01,1/2,sqrt(1/(4*79080)),lower.tail=F)
[1] 0.5041363
```

Ce seuil est supérieur car on veut prendre moins de risque de se tromper, c'est-à-dire qu'on veut moins "rater" de valeurs d'échantillon en accord avec l'hypothèse. En revanche il est évident qu'on augmente le risque de seconde espèce dans ce cas (on accepte plus de valeurs pouvant provenir d'un échantillon qui suit la loi sous H_1).

2. *La proportion de garçons est de 105/205.*

Hypothèses

Ici l'hypothèse à tester est H_0 : "$p = \frac{105}{205}$". On réalise ici un test bilatéral, à savoir que l'hypothèse alternative que l'on teste est H_1 : "$p \neq \frac{105}{205}$".
Sous l'hypothèse H_0, cela signifie donc que le nombre de naissances de garçons dans notre échantillon suit une loi $\mathcal{N}\left(frac105205, \sqrt{\frac{\frac{105}{205}\left(1-\frac{105}{205}\right)}{79080}}\right)$.

Zone critique
On veut que la probabilité que le nombre de naissances de garçons soit en-dehors de l'intervalle critique ne dépasse pas un certain risque, lorsque Y suit la loi sous l'hypothèse H_0. Par exemple, prenons un risque de 5%. On veut connaître les bornes t_1 et t_2 telles que $P(t_1 > Y$ ou $Y > t_2) = 5\%$. Autrement dit, on veut connaître t_1 et t_2 tels que : $P(t_1 \leq Y \leq t_2) = 0.95$. On normalise la variable aléatoire pour se ramener à une loi normale centrée réduite : $P(\frac{t_1-p}{\sqrt{\frac{p(1-p)}{n}}} < U < \frac{t_2-p}{\sqrt{\frac{p(1-p)}{n}}}) = 0.95$ où U suit une loi normale centrée réduite. Or l'intervalle critique est centré autour de la moyenne de la loi sous H_0, donc on a un certain u tel que :

$$\begin{cases} -u & = & \frac{t_1-p}{\sqrt{\frac{p(1-p)}{n}}} \\ u & = & \frac{t_2-p}{\sqrt{\frac{p(1-p)}{n}}} \end{cases}$$

Ce u vérifie : $P(-u < U < u) = 0.95$, soit, comme déjà vu dans le TD1, $P(U < u) = \frac{1+0.95}{2}$.
Application numérique :

```
> u <- qnorm(1.95/2)
> u
[1] 1.959964
> t1 <- -u*sqrt((105/205)*(1-105/205)/79080) + 105/205
> t2 <- u*sqrt((105/205)*(1-105/205)/79080) + 105/205
> t1
[1] 0.5087113
> t2
[1] 0.5156789
```

L'intervalle critique est alors $[0.5087113, 0.5156789]$.

Décision Si la valeur prise par l'échantillon est dans l'intervalle, on accepte H_0, sinon on rejette H_0. Ici la valeur prise par l'échantillon est 0.5106222, c'est dans l'intervalle, donc on accepte H_0 (au risque de 5%).

3 Le problème « Effet de tailles »

On jette une pièce de monnaie et on compte le nombre de fois où l'on obtient « face ». On considère les trois cas suivants :

— On jette 100 fois la pièce et on obtient 55 fois « Face ».
— On jette 1000 fois la pièce et on obtient 550 fois « Face ».
— On jette 10000 fois la pièce et on obtient 5500 fois « Face ».

Exercice 15 *Dans chacun de ces cas, peut-on accepter l'hypothèse selon laquelle la probabilité d'obtenir « Face » est 0,5 ?*

On va réaliser l'étude théorique avec un échantillon de taille n, puis on passera à l'application numérique.

Hypothèses
On modélise les lancers de pièce par une variable aléatoire suivant une loi de Bernoulli $\mathcal{B}e(p)$. Le nombre de "face" obtenu en n lancers suit alors une loi binomiale à n expérience et de paramètre p. On peut l'approximer par une loi normale $\mathcal{N}(np, \sqrt{np(1-p)})$.

L'hypothèse à tester est H_0 : "$p = \frac{1}{2}$". Vu les résultats de l'échantillon, on peut faire un test unilatéral dont l'alternative est donc H_1 : "$p > \frac{1}{2}$".

Zone critique
On veut que la probabilité que le nombre de "face" soit supérieur à la valeur critique ne dépasse pas un certain risque, lorsque Y suit la loi sous l'hypothèse H_0. Par exemple, prenons un risque de 5%. On veut connaître le seuil s tel que $P(Y > s) = 5\%$. On normalise la variable aléatoire puis on lit dans les tables la valeur de $u = \frac{s-np}{\sqrt{np(1-p)}}$.

```
> qnorm(0.05,lower.tail=F)
[1] 1.644854
```

$$s = \sqrt{np(1-p)} \times 1.644854 + np$$

Décision
Si la valeur prise par l'échantillon est inférieure au seuil, on accepte H_0, sinon on rejette H_0.

Application numérique

1. $n = 100$:

   ```
   > u <- qnorm(0.05,lower.tail=F)
   > s <- sqrt(100*0.5*0.5)*u + 100*0.5
   > s
   [1] 58.22427
   ```

   La valeur pour l'échantillon est 55, donc inférieure au seuil : on accepte H_0.

2. $n = 1000$:

   ```
   > s <- sqrt(1000*0.5*0.5)*u + 1000*0.5
   > s
   [1] 526.0074
   ```

   La valeur pour l'échantillon est 550, donc supérieure au seuil : on rejette H_0.

3. $n = 10000$:

```
> s <- sqrt(10000*0.5*0.5)*u + 10000*0.5
> s
[1] 5082.243
```

La valeur pour l'échantillon est 5500, donc (très) supérieure au seuil : on rejette H_0.

La décision est différente selon les cas, alors qu'on aurait pu s'attendre au contraire. En effet l'écart relatif entre la valeur théorique et la valeur de l'échantillon est toujours 10%, elle est donc proportionnelle à la valeur théorique. Mais si on regarde la variation de l'écart-type en fonction de n, on s'aperçoit qu'il a une croissance en $\sqrt{n}$, donc asymptotiquement négligeable devant n. Ainsi, l'écart observé "grandit" beaucoup plus vite que l'écart "autorisé" (l'amplitude de l'intervalle critique est proportionnelle à l'écart-type).