Chapitre I : Introduction aux tests statistiques

Cours de tests paramériques

2024-2025

Plan du cours

- Chapitre I : Introduction aux tests statistiques
- Chapitre II : Tests paramétriques à un échantillon
- Chapitre III : Tests de comparaison d'échantillons
- Chapitre IV : Tests du Chi-deux d'indépendance et d'homogénéité
- Chapitre V : Coefficients de corrélation et tests

Planning et évaluations

- 11 séances de cours d'1h30
- 8 séances de TD d'1h30 en 1/2 groupe
- 3 séances de TP d'1h30 sous `R` en 1/4 de promo
- 2 DS d'1h30 les (10 ou 15) Octobre* et 19 Décembre*
- 1 compte-rendu de TP en binôme (20/12)

$\longrightarrow$ 3 notes coefficientées 2+2+1

Programme National

- Apprentissages critiques :
  - *AC22.04* : Appréhender l'idée de confronter une hypothèse avec la réalité pour prendre une décision
  - *AC22.05* : Apprécier les limites de validité et les conditions d'application d'une analyse
- SAE associée : *SAÉ 3.EMS.01* : Recueil et analyse de données par échantillonnage ou plan d'expérience

Introduction aux tests : Exemple 1 - Curly

Un producteur affirme qu'il y a en moyenne 150 cacahuètes dans un paquet classique de *Curly*. Un étudiant pointilleux se charge de vérifier cette déclaration. Pour cela, il sélectionne aléatoirement 10 paquets de *Curly* au supermarché et compte manuellement le nombre de cacahuètes que ceux-ci contiennent. Il obtient, en plus d'une indigestion, les observations suivantes :

```
x <- c(153,142,160,148,147,151,142,156,149,140)
mean(x)
```

```
## [1] 148.8
```

Peut-on affirmer que le producteur est un voleur?

Introduction aux tests : Exemple 1 - Curly

Un producteur affirme qu'il y a en moyenne 150 cacahuètes dans un paquet classique de *Curly*. Un étudiant pointilleux se charge de vérifier cette déclaration. Pour cela, il sélectionne aléatoirement 10 paquets de *Curly* au supermarché et compte manuellement le nombre de cacahuètes que ceux-ci contiennent. Il obtient, en plus d'une indigestion, les observations suivantes :

```
x <- c(153,142,160,148,147,151,142,156,149,140)
mean(x)
```

```
## [1] 148.8
```

Peut-on affirmer que le producteur est un voleur?

- Même si, sur cet échantillon de 10 paquets, la moyenne est < 150 cacahuètes, est-ce que cela suffit pour accuser le producteur de tromperie?
- Si, en prenant un autre échantillon, on avait trouvé 158 cacahuètes en moyenne, qu'aurait-on pu en conclure?
- Aurait-il été plus facile de conclure en prenant un échantillon de 100 paquets?

Introduction aux tests : Exemple 2 - Alcool et contrôle

Un prof de maths décide, sans le savoir, de faire son examen le lendemain d'un repas de classe très arrosé. Après avoir corrigé les copies, il obtient les notes suivantes :

```
##  [1] 17  8  2  9 16 12  6  1  6 10 15 15 11  9  4 10 17  2  7 12
## [26] 16 14  9 17  2
```

```
## pour une moyenne de 9.9.
```

Très déçu, il décide de comparer ces notes avec celles obtenues par une autre classe B de même niveau qui a eu le même examen la veille.

```
##  [1] 18 12 19  3  3 16  2 12  2 15  7 12 15 16  3 10 19  1 15  5
## [26] 15 13  8 13  0
```

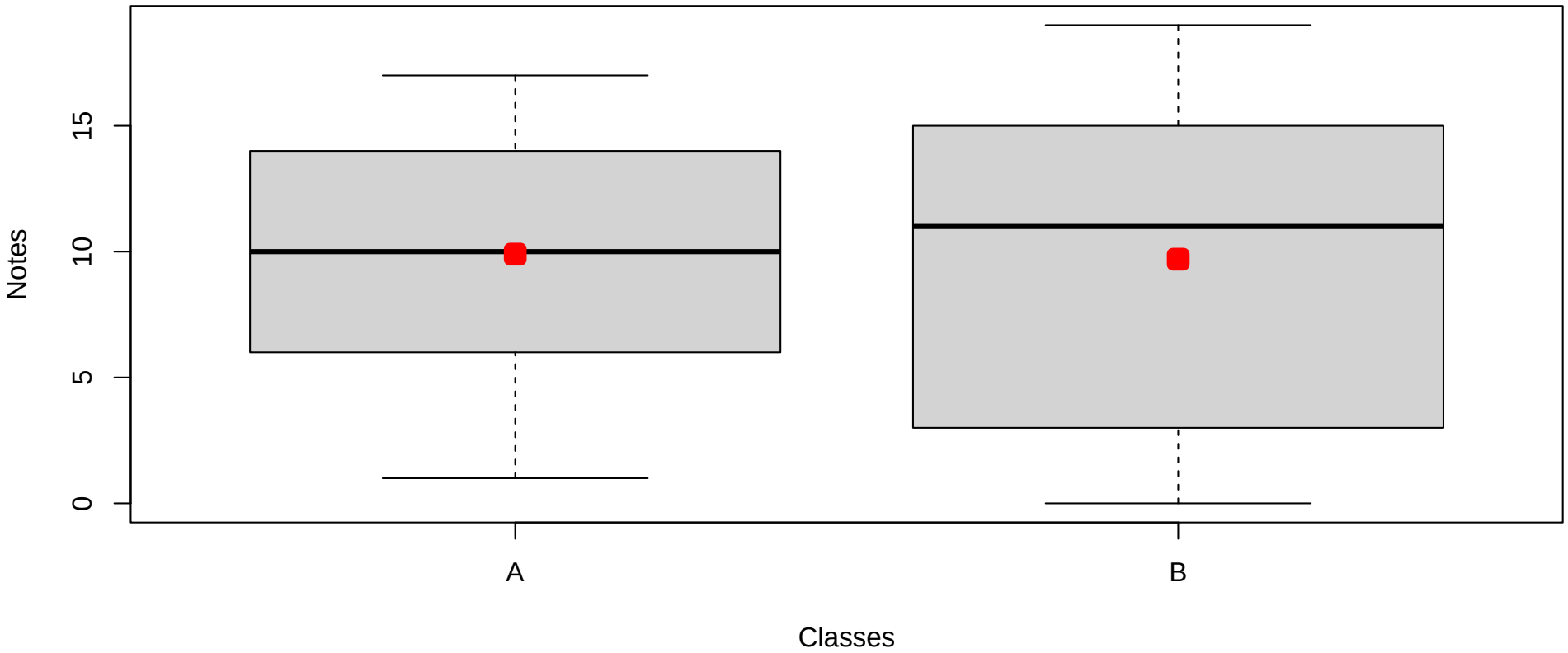
```
## pour une moyenne de 9.7.
```

Introduction aux tests : Exemple 2 - Alcool et contrôle

En comparant les moyennes, peut-on en conclure qu'il vaut mieux sortir avant un contrôle de maths?

Introduction aux tests : Exemple 2 - Alcool et contrôle

En comparant les moyennes, peut-on en conclure qu'il vaut mieux sortir avant un contrôle de maths?



La conclusion semble moins évidente!

Introduction aux tests : Exemple 3 - Cigarette

Dans le cadre d'une enquête sur le comportement des jeunes vis-à-vis de la cigarette, on sonde un panel de lycéens. Les observations suivantes ont été obtenues :

```
##         Cigarette
## Sexe     Fumeur Non fumeur
##   Femme  0.160       0.840
##   Homme  0.335       0.665
```

- Existe t'il un lien entre le tabagisme et le sexe chez les jeunes?

Introduction aux tests : Exemple 3 - Cigarette

Dans le cadre d'une enquête sur le comportement des jeunes vis-à-vis de la cigarette, on sonde un panel de lycéens. Les observations suivantes ont été obtenues :

```
##         Cigarette
## Sexe     Fumeur Non fumeur
##   Femme  0.160       0.840
##   Homme  0.335       0.665
```

- Existe t'il un lien entre le tabagisme et le sexe chez les jeunes?

```
##         Cigarette
## Sexe         Fumeur Non fumeur
##   Femme 0.1066667  0.2400000
##   Homme 0.8933333  0.7600000
```

- Cette différence est-elle dûe à une différence de proportion liée à l'échantillon ou est-elle exportable à la population entière?

Introduction aux tests

A partir de ces trois exemples, on voit que les résultats observés dépendent :

- de la taille de l'échantillon : on a l'intuition que plus l'échantillon est grand, plus il sera facile de discriminer entre 2 hypothèses,
- de la variance des observations : on a l'intuition que plus la variance est petite, plus il sera facile de discriminer entre 2 hypothèses,
- des échantillons choisis : en pratique, on ne dispose que d'un échantillon, est-ce que celui-ci est bien représentatif de la population? Ou est-ce un échantillon atypique?

Tests statistiques : Hypothèses

Objectif : valider ou non une hypothèse faite sur une ou plusieurs populations.

En pratique, on dispose de deux hypothèses, l'hypothèse **nulle** (H_0) et l'hypothèse **alternative** (H_1) et le but est d'essayer de discriminer entre ces deux hypothèses.

Dans l'exemple sur les *Curly*, on représente par X la variable aléatoire (v.a) "nombre de cacahuètes dans un paquet classique de Curly' '. On note $\mu = \mathbb{E}[X]$ et $\sigma^2 = \mathbb{V}[X]$. On teste alors :

$$(H_0) \ : \ \mu = 150 \text{ contre } (H_1) \ : \ \mu \neq 150$$

Il existe deux *mondes* :

- dans le *monde* de (H_0), un paquet de Curly contient 150 cacahuètes en moyenne, il n'y aucune arnaque de la part du producteur.
- dans le *monde alternatif* de (H_1), un paquet de Curly ne contient pas 150 cacahuètes en moyenne. Si c'est plus, tant mieux, sinon ça sent l'arnaque!

Tests statistiques : Hypothèses

Ceci peut avoir des conséquences plus dramatiques : supposons que plusieurs personnes aient soufferts d'intoxication alimentaire de salmonelle. Après enquête, on suspecte une marque de glaces d'en être responsable. On note X la v.a "quantité de salmonelle dans un pot de glace' ' et $\mu = \mathbb{E}[X]$. Le niveau réglementaire que doit respecter un fabricant de glaces étant de 0.3 NPP/g, on veut tester :

$$(H_0) : \mu \leq 0.3 \text{ contre } (H_1) : \mu > 0.3$$

- Dans le *monde* de (H_0), la quantité de salmonelle des pots de glace ne dépasse pas le niveau réglementaire de 0.3 NPP/g.
- Dans le *monde alternatif* de (H_1), la quantité de salmonelle des pots de glace dépasse le niveau réglementaire de 0.3 NPP/g et une personne peut tomber **gravement** malade si elle mange de la glace!

Tests statistiques : Risques

A chacune des deux hypothèses est associée un risque :

- dans le monde de (H_0) (les pots de glace ne présentent pas de risque d'intoxication), on peut se **tromper** en choisissant (H_1) et en disant que la glace peut être dangereuse pour la santé (alors que ce n'est pas le cas)!

Ce risque s'appelle le risque de **1ère espèce**.

- dans le monde de (H_1) (les pots de glace sont dangereux pour la santé), on peut se **tromper** en choisissant (H_0) à la place et en disant que la glace n'est pas dangereuse pour la santé (alors qu'elle l'est)!

Ce risque s'appelle le risque de **2ème espèce**.

Tests statistiques : Risques

On voit que les deux risques ne sont pas symétriques! Selon vous, lequel est le plus **grave**?

Tests statistiques : Risques

On voit que les deux risques ne sont pas symétriques! Selon vous, lequel est le plus **grave**?

Cela dépend du point de vue :

- pour le fabricant peu scrupuleux, le plus **grave** est de dire que sa glace peut être dangereuse pour la santé alors qu'elle ne l'est pas!
- pour le consommateur, le plus **grave** est de dire que la glace n'est pas dangereuse pour la santé alors qu'elle l'est!

En pratique on n'arrive généralement pas à contrôler les deux risques. Il faut choisir ce que l'on veut montrer et quel risque on veut contrôler.

Tests statistiques : Risques

Si on reprend l'exemple précédent de l'intoxication alimentaire où l'on teste :

$$(H_0) : \mu \leq 0.3 \text{ contre } (H_1) : \mu > 0.3$$

	Décision	
Réalité	(H_0)	(H_1)
(H_0)	$1-\alpha$	α
(H_1)	β	$1-\beta$

- α s'appelle le risque de **1ère espèce**.

$$\alpha = \mathbb{P}_{H_0}[\text{rejeter } H_0].$$

- β s'appelle le risque de **2ème espèce**.

$$\beta = \mathbb{P}_{H_1}[\text{ne pas rejeter } H_0].$$

Tests statistiques : Risques

En pratique, le risque de **1ère espèce** α est fixé à l'avance $(5\%, 10\%)$. C'est le risque que l'on contrôle :

- si on décide de rejeter (H_0), on connaît alors le pourcentage d'erreur de se tromper $(5\%, 10\%)$,
- si on décide de ne pas rejeter (H_0), on ne connaît par contre généralement pas la marge d'erreur que l'on commettrait en choisissant (H_0).

Astuce : puisqu'on connaît le pourcentage d'erreur de se tromper en rejetant (H_0) sous l'hyptohèse (H_0), (H_1) est toujours l'hypothèse que l'on veut montrer.

Tests statistiques : Règle de décision

Pour vérifier si la marque de glaces est bien responsable de l'intoxication alimentaire, les services sanitaires réquisitionnent de manière aléatoire un échantillon de pots de glace de taille n ($n = 9$) et mesurent la quantité de salmonelle X qu'ils contiennent. On note $\mu := \mathbb{E}[X]$. Les services sanitaires étant débutants en tests statistiques, ils simplifient les hypothèses sous la forme :

$$(H_0) : \mu = 0.3 \text{ contre } (H_1) : \mu = 0.4$$

A partir de quelle quantité moyenne de salmonelle observée va t'on rejeter (H_0)? Et avec quel risque?

Tests statistiques : Règle de décision

Pour vérifier si la marque de glaces est bien responsable de l'intoxication alimentaire, les services sanitaires réquisitionnent de manière aléatoire un échantillon de pots de glace de taille n ($n = 9$) et mesurent la quantité de salmonelle X qu'ils contiennent. On note $\mu := \mathbb{E}[X]$. Les services sanitaires étant débutants en tests statistiques, ils simplifient les hypothèses sous la forme :

$$(H_0) \ : \ \mu = 0.3 \text{ contre } (H_1) \ : \ \mu = 0.4$$

A partir de quelle quantité moyenne de salmonelle observée va t'on rejeter (H_0)? Et avec quel risque?

Le risque de 1ère espèce α étant fixé, il faut établir une **règle de décision**, c'est-à-dire un **seuil critique** pour μ à partir duquel on choisit de rejeter (H_0) avec risque α.

Tests statistiques : Règle de décision

Si on formalise les choses :

- on note X la v.a "quantité de salmonelle contenue dans un pot de glace' ',
- on suppose que X suit une *loi gaussienne* d'espérance μ et de variance connue égale à 0.08 :

$$X \sim \mathcal{N}(\mu, 0.08),$$

- on considère $X_1,\ldots,X_n$ n v.a de même loi que X,
- on dispose de $n = 9$ réalisations $x_1,\ldots,x_n$ de $X_1,\ldots,X_n$ (valeurs observées sur l'échantillon de pots de glace par les services sanitaires).

La variable d'intérêt est ici définie par :

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

qui est un bon estimateur du paramètre inconnu μ.

Tests statistiques : Règle de décision

Rappel :

$$\forall X \sim \mathcal{N}(\mu, \sigma^2), \quad \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1) \quad \text{donc} \quad \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1).$$

Rappelons les hypothèses :

$$(H_0) \ : \ \mu = 0.3 \text{ contre } (H_1) \ : \ \mu = 0.4$$

- Sous (H_0), $\sqrt{9}(\bar{X}_n - 0.3)/\sqrt{0.08} \sim \mathcal{N}(0,1)$.
- Sous (H_1), $\sqrt{9}(\bar{X}_n - 0.4)/\sqrt{0.08} \sim \mathcal{N}(0,1)$.

On définit la variable $T_n = \sqrt{9}(\bar{X}_n - 0.3)/\sqrt{0.08}$ comme la **statistique de test**.

Astuce : La loi de T_n sous (H_0) doit être connue pour nous permettre de comparer la valeur observée t_n de T_n avec ce qui est attendu!

Tests statistiques : Règle de décision

Objectif : définir un seuil à partir duquel on choisit (H_1) plutôt que (H_0) en contrôlant la marge d'erreur.

Soit c_α un seuil à définir à risque α fixé. La **règle de décision** est la suivante :

- si $T_n > c_\alpha$, on rejette (H_0),
- si $T_n \leq c_\alpha$, on ne rejette pas (H_0).

La loi de T_n étant connue $(\mathcal{N}(0,1))$, le seuil c_α est trouvé sur la table de la loi normale. Si $\alpha = 0.05$, il vaut $c_\alpha = 1.645$ et nous assure alors que :

$$\mathbb{P}_{H_0}[\text{rejeter } H_0] = \mathbb{P}_{H_0}[T_n > c_\alpha] = \alpha.$$

Si, sur l'échantillon prélevé, on a mesuré $\bar{x}_n = 0.38$, alors :

$$t_n = \sqrt{9}(0.38 - 0.3)/\sqrt{0.08} \approx 0.849.$$

Conclusion :

Tests statistiques : Règle de décision

Objectif : définir un seuil à partir duquel on choisit (H_1) plutôt que (H_0) en contrôlant la marge d'erreur.

Soit c_α un seuil à définir à risque α fixé. La **règle de décision** est la suivante :

- si $T_n > c_\alpha$, on rejette (H_0),
- si $T_n \leq c_\alpha$, on ne rejette pas (H_0).

La loi de T_n étant connue ($\mathcal{N}(0,1)$), le seuil c_α est trouvé sur la table de la loi normale. Si $\alpha = 0.05$, il vaut $c_\alpha = 1.645$ et nous assure alors que :

$$\mathbb{P}_{H_0}[\text{rejeter } H_0] = \mathbb{P}_{H_0}[T_n > c_\alpha] = \alpha.$$

Si, sur l'échantillon prélevé, on a mesuré $\bar{x}_n = 0.38$, alors :

$$t_n = \sqrt{9}(0.38 - 0.3)/\sqrt{0.08} \approx 0.849.$$

Conclusion : on ne rejette pas (H_0). Le test n'est pas significatif au niveau 5%. On a pas mis en evidence que les glaces avaient trop de salmonelles.

Tests statistiques : Degré de significativité ou p-valeur

Si on ne rejette pas (H_0) au risque 5%, rejetterait-on (H_0) au risque 15%?

Tests statistiques : Degré de significativité ou p-valeur

Si on ne rejette pas (H_0) au risque 5%, rejetterait-on (H_0) au risque 15%?

Effectivement, en augmentant le risque de se tromper, on a plus de chance de pouvoir rejetter (H_0). Un indicateur intéressant est le **degré de significativité** ou **p-valeur** du test qui représente le plus petit seuil pour lequel on rejette (H_0). Il est défini par :

$$\text{p-valeur} = \mathbb{P}_{H_0}[T_n > t_n],$$

c'est-à-dire la probabilité d'avoir observé une valeur aussi extrême pour T_n que celle observée t_n.

Ici, p-valeur $= \mathbb{P}_{H_0}[T_n > 0.849] = \ldots = 0.198$. Cela signifie donc que si (H_0) était vraie (les pots de glace ne contenaient pas trop de salmonelle), on aurait alors environ 19.8% de chances de trouver autant de salmonelle que celle observée dans ces 9 pots. Ce n'est pas suffisant pour conclure.

Tests statistiques : Risque de deuxième espèce

On peut aussi calculer le risque de 2ème espèce : le risque que l'on prend en choisissant (H_0) à la place de (H_1), c'est-à-dire, en déclarant que les glaces ne sont pas dangereuses à tort!

$$\begin{aligned} \beta &= \mathbb{P}_{H_1}[\text{ne pas rejeter } H_0] \\ &= \mathbb{P}_{H_1}[T_n \leq c_\alpha] \\ &= \mathbb{P}_{H_1}[T_n \leq 1.645] = \ldots = 0.719. \end{aligned}$$

On a donc environ 71.9% de chances de se tromper en choisissant (H_0) à la place de (H_1).

Tests statistiques : Puissance de test

La **puissance** d'un test représente la probabilité de prendre la bonne décision en rejettant l'hypothèse nulle :

$$\pi = \mathbb{P}_{H_1}[\text{rejeter } H_0] = 1 - \beta.$$

- La puissance d'un test est d'autant plus grande que le risque de 2ème espèce est petit.
- La puissance d'un test est d'autant plus grande que l'hypothèse alternative (H_1) est éloignée de l'hypothèse nulle (H_0).

Tests statistiques : Influence des paramètres

Niveau du test

Le risque de 1ère espèce α est aussi appelé **niveau** du test. Que se passe-t'il quand le niveau du test augmente?

Tests statistiques : Influence des paramètres

Niveau du test

Le risque de 1ère espèce α est aussi appelé **niveau** du test. Que se passe-t'il quand le niveau du test augmente?

- on tolère une plus grande probabilité d'erreur en rejetant (H_0) alors qu'elle était vraie,
- la règle de décision est moins stricte : on rejette plus souvent (H_0),
- l'erreur de 2ème espèce β diminue,
- la puissance π augmente.

Et inversement quand le niveau du test diminue.

Objectif : mettre en place un test avec niveau de test le plus petit possible et une puissance la plus grande possible. Ceci peut être fait en augmentant la taille de l'échantillon.

Tests statistiques : Influence des paramètres

Formulation des hypothèses

Que se passe t'il lorsque l'on échange les hypothèses du test?

$$(H_0') : \mu = 0.4 \text{ contre } (H_1') : \mu = 0.3$$

Tests statistiques : Influence des paramètres

Formulation des hypothèses

Que se passe t'il lorsque l'on échange les hypothèses du test?

$$(H_0') : \mu = 0.4 \text{ contre } (H_1') : \mu = 0.3$$

La statistique de test devient :

$$T_n = \sqrt{9}(\bar{X}_n - 0.4)/\sqrt{0.08} \sim_{H_0'} \mathcal{N}(0,1).$$

Tests statistiques : Influence des paramètres

Formulation des hypothèses

Que se passe t'il lorsque l'on échange les hypothèses du test?

$$(H_0') : \mu = 0.4 \text{ contre } (H_1') : \mu = 0.3$$

La statistique de test devient :

$$T_n = \sqrt{9}(\bar{X}_n - 0.4)/\sqrt{0.08} \sim_{H_0'} \mathcal{N}(0,1).$$

- A $\alpha = 0.05$ fixé, le seuil critique devient $c_\alpha = -1.645$,
- La zone de rejet pour H_0' devient $R_{0.05} = \{T_n < 1.645\}$,
- La valeur observée de T_n devient $t_n = \sqrt{9}(\bar{x}_n - 0.4)\sqrt{0.08} \approx -0.21$,
- La conclusion : on ne rejette pas $(H_0)'$, Si on part du principe que les glaces ont la salmonelle on a pas mis en evidence qu'elles ne l'avaient pas.
- La p-valeur et la puissance sont recalculées et valent 0.5832 et 0.0035.

Tests statistiques : Application sous R

Sous R, la procédure de test est très facile à implémenter :

```
x <- c(0.175,0.205,0.76,0.719,0.199,0.529,0.306,0.52,0.01)
t.test(x=x,alternative = "greater",mu=0.3)
```

```
##
##  One Sample t-test
##
## data:  x
## t = 0.92005, df = 8, p-value = 0.1922
## alternative hypothesis: true mean is greater than 0.3
## 95 percent confidence interval:
##  0.2179689       Inf
## sample estimates:
## mean of x
## 0.3803333
```