# Estimation

Université de Bourgogne, Licence 2 Gestion

October 2, 2024

# Introduction à la stat. inférentielle

Considérons une population et une variable d'intérêt $X$ sur cette population

- Un *échantillon de taille n* est la donnée des valeurs de $X$ pour $n$ individus tirés au hasard dans la population.
- Le but de la *statistique inférentielle* est de donner des résultats sur la variable $X$ (moyenne, écart-type...) en se basant sur un échantillon.
- et en quantifiant la probabilité de faire une erreur.
- Elle s'appuie sur l'utilisation des probabilités **en considérant l'échantillon comme les valeurs de variables aléatoires indépendantes de même loi que** $X$.
- Deux méthodes : l'estimation et les tests d'hypothèses.
- On va traiter ici l'estimation d'une proportion, d'une moyenne et d'une variance (ou d'un écart-type)

# Estimation d'une proportion

- Une usine affirme produire 90% de pièces aux normes. Pour contrôler ceci, une association de consommateurs fait prélever un échantillon de 200 pièces et trouve 172 pièces aux normes.
- C'est-à-dire 86%.
- Quelle conclusion peut-elle donner? avec quel risque d'erreur?

# Modélisation

- Dès qu'il est question de proportion, la v.a. est une Bernoulli.
- Ici c'est la v.a. $X$ qui vaut
  - 1 si la pièce contrôlée est aux normes.
  - 0 sinon.
- Notons $p$ la probabilité du 1.
- On peut voir $p$ comme "la proportion de pièces aux normes dans la population (infinie) des pièces produites par l'usine" : l'usine prétend que $p = 0.90$.
- L'échantillon est une suite de 1 et de 0 de taille 200 :
  1 1 0 1 0 1 1 1 1 0 1 1 0 0 1 1 1 ...
- avec 172 fois le chiffre 1.
- On considère ces 1 et ces 0 comme les valeurs prises par des variables aléatoires $X_1, ..., X_{200}$ indépendantes de loi $B(p)$.

- La proportion de "1" dans un échantillon vaut

$$P_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

- $P_n$ est une variable aléatoire (dépendant du tirage de l'échantillon) appelée **proportion aléatoire** de pièces aux normes.
- Vocabulaire : on dit que $P_n$ est une statistique car c'est une v.a. qui se calcule sur des échantillons.
- Dans cet échantillon particulier, la valeur de la statistique $P_n$ est $p^e = 0.86$.

# Attention, il ne faut pas confondre

- $p$ =proportion de pièces aux normes dans la population
  $p$ est inconnue.
- $p^e$ = proportion de pièces aux normes dans l'échantillon
  $p^e = 0.86$.
- $P_n$ = proportion aléatoire de pièces aux normes
  $P_n$ est une statistique.

# Echantillonnage de la proportion

- Imaginons qu'un autre échantillon de 200 pièces donne $p_1^e = 0.84 \neq p^e$.
- Le lien entre $p^e$ et $p_1^e$ est :
  ce sont deux réalisations de la même statistique $P_n$.
- Imaginons qu'on prélève 100 échantillons de 200 pièces chacun :
- on obtient ainsi 100 proportions expérimentales :
  $p_1^e, p_2^e, ..., p_{100}^e$.
- Imaginons que $p = 0.90$, comme l'affirme l'usine. On observerait alors que ces valeurs observées ont tendance à osciller autour de 0.90.
- Plus précisément : l'histogramme de ces 100 valeurs aurait l'allure d'une loi normale de moyenne 0.90.
- Plus généralement :

# Théorème de Moivre-Laplace (autre version)

☞ si $n \geq 30$ , $np > 5$, $n(1-p) > 5$ :

$P_n$ peut être approximée par une loi normale $\mathcal{N}\left(p; \frac{p(1-p)}{n}\right)$.

☞ Ce résultat s'appelle *théorème d'échantillonnage pour une proportion*. Il sert aux intervalles de confiance et aux tests pour une proportion.

# Estimation de $p$

- L'estimation ponctuelle de $p$ est $p^e = 0.86$.
- Pour obtenir un intervalle de confiance, on introduit la statistique

$$Z = \frac{P_n - p}{\sqrt{\frac{p(1-p)}{n}}}$$

- D'après le résultat, $Z$ suit une loi normale centrée réduite.
- Si on fixe un petit nombre $\alpha$ (risque d'erreur), on sait trouver un quantile $z_\alpha$ tel que

$$\mathbb{P}(-z_\alpha < Z < z_\alpha) = 1 - \alpha$$

- $1 - \alpha$ est appelé la confiance.

# Intervalle de confiance de $p$

- On déduit $\mathbb{P}(-z_\alpha < \frac{P_n - p}{\sqrt{\frac{p(1-p)}{n}}} < z_\alpha) = 1 - \alpha$
- On isole la proportion inconnue $p$ :

$$\mathbb{P}\left(P_n - z_\alpha\sqrt{\frac{p(1-p)}{n}} < p < P_n + z_\alpha\sqrt{\frac{p(1-p)}{n}}\right) = 1 - \alpha$$

- En posant $a_\alpha = z_\alpha\sqrt{\frac{p(1-p)}{n}}$, on voit que :

$$\mathbb{P}(P_n - a_\alpha < p < P_n + a_\alpha) = 1 - \alpha$$

- ☞ C'est-à-dire que $(1-\alpha)\%$ des intervalles $[P_n - a_\alpha, P_n + a_\alpha]$ contiennent $p$.
- ☞ Etant donné un échantillon de proportion expérimentale $p^e$, on appelle intervalle de confiance $1-\alpha$ pour $p$ l'intervalle:

$$\mathcal{I}_{1-\alpha}(p) = [p^e - a_\alpha, p^e + a_\alpha]$$

où $a_\alpha = z_\alpha \sqrt{\frac{p^e(1-p^e)}{n}}$

# Retour sur l'exemple

- Suivons la procédure décrite dans le formulaire:
  - Fixons par exemple une confiance à $1-\alpha = 95\%$ (donc un risque d'erreur à $\alpha = 5\% = 0.05$)
  - On cherche $z_\alpha$ tel que $\mathbb{P}(Z > z_\alpha) = 0.05/2 = 0.025$
  - $z_\alpha = 1.96$
  - donc $a_\alpha = 1.96\sqrt{\frac{0.86\times(1-0.86)}{200}} = 0.048$
  - L'intervalle de confiance 95 % pour $p$ est donc

$$\mathcal{I}_{95\%}(p) = [0.86 - 0.048, 0.86 + 0.048] = [0.812, 0.908]$$

- ☞ Ceci signifie que *avec une confiance de 95 %, on peut dire que la proportion de pièces aux normes produites par l'usine est entre 81.2 % et 90.8 %*
- ☞ **Conclusion** L'association de consommateurs ne peut pas dire que l'affirmation de l'usine ($p = 90\%$) est fausse.

# Retour sur la signification de la confiance

- Fixons une confiance : 95% par exemple.
- Par la procédure décrite, à chaque échantillon est associé un intervalle de confiance 95% pour $p$.
- Ceci signifie que :
  95% **des échantillons donnent un intervalle contenant** $p$.
- On peut voir la confiance comme
  la confiance que l'on a en notre procédure d'estimation

# Exercice 9

- Dans une entreprise de matériel informatique, on veut estimer la proportion $p$ d'ordinateurs dont la durée de vie est inférieure à 4 ans. On a relevé un échantillon de 240 ordinateurs livrés à des clients depuis 4 ans. On a constaté qu'il y en a 96 en fin de vie. Donner une estimation de $p$ à l'aide d'un intervalle de confiance à 90 %.
- Quelle taille minimale d'échantillon faudrait-il avoir pour estimer cette proportion à 0.03 près avec un intervalle de confiance à 90 % ?.

# Cas particulier d'une population finie

- Si la taille de la population est finie et connue, on a intérêt à en tenir compte pour calculer l'intervalle de confiance.
- Notons $N$ la taille de la population et toujours $n$ la taille de l'échantillon.
- Le rayon de l'intervalle de confiance devient alors :

$$a_\alpha = z_\alpha \sqrt{\frac{p^e(1-p^e)}{n}} \times \sqrt{1-\frac{n}{N}}$$

- Comme $\sqrt{1-\frac{n}{N}} < 1$, ce rayon est plus petit donc l'intervalle de confiance est plus précis.

# Exercice 10

On s'intéresse à l'estimation de la proportion $p$ d'individus atteints par une maladie professionnelle dans une entreprise de 1500 salariés. On sait par ailleurs que trois personnes sur dix sont ordinairement touchées par cette maladie dans des entreprises du même type.

☞ Quelle taille d'échantillon faut-il sélectionner pour avoir une estimation à 0.03 près avec une confiance à 95% dans le cas :

- sans tenir compte de l'information sur la taille de l'entreprise.
- en en tenant compte.

# Estimation d'une moyenne

- Soit $X$ une variable quantitative prenant les valeurs $x_1, ..., x_n$ sur un échantillon de $n$ individus issu d'une population $\mathcal{P}$.
- Le paramètre que l'on veut estimer est $\mu = \mathbb{E}(X)$, "moyenne" de $X$ sur la population.
- La moyenne de cet échantillon est

$$m^e = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

(appelée **moyenne expérimentale**)

- $m^e$ est une estimation ponctuelle de $\mu$.

- Soient $X_1, ..., X_n$ les variables aléatoire indépendantes définissant l'échantillon aléatoire.
- Considérons la statistique :

$$\overline{X_n} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

(appelée **moyenne aléatoire**).

- Considérons également la statistique $S_n$ qui donne l'écart-type de l'échantillon:

$$S_n = \sqrt{\frac{1}{n} \sum_{i=1}^{n} X_i^2 - \overline{X_n}^2}$$

(appelée **écart-type aléatoire**)

# Echantillonnage de la moyenne

- Selon l'échantillon prélevé, la réalisation observée de la moyenne aléatoire $\overline{X_n}$ bouge. Mais COMMENT?
- La loi qui intervient ici est la loi Student. Pour obtenir cette loi, il faut introduire une autre statistique :

$$T = \sqrt{n-1}\frac{\overline{X_n} - \mu}{S_n}$$

- **Théorème d'échantillonnage de la moyenne** *La statistique $T$ suit une loi de Student à $n-1$ degrés de liberté.*
- Rem : pour $n \leq 30$, c'est vrai uniquement si X suit une loi normale.
- Ce théorème permet de calculer l'intervalle de confiance en utilisant la procédure décrite dans le formulaire.

# Exercice 7

- ☞ Dans le cadre du lancement d'un nouveau produit, une entreprise a mené une enquête de satisfaction client. La satisfaction est mesurée sur une échelle de 1 à 100. Les études préliminaires montrent que la satisfaction moyenne attendue est de 75, avec un écart-type de 10.
- ☞ Sachant cela, est-il possible d'observer une moyenne de satisfaction inférieure à 73 sur un échantillon de 80 clients?

# Exercice 8

☞ 1) On a relevé le nombre d'heures (x) travaillées par mois pour un échantillon de 100 employés d'un certain type d'usine dans le but d'estimer le nombre moyen d'heures travaillées dans la population des employés de ce type d'usine.

| x | [141;143[ | [143;145[ | [145;147[ | [147;149[ | [149;151[ |
|---|---|---|---|---|---|
| eff. | 1 | 5 | 6 | 21 | 32 |

| x | [151;153[ | [153;155[ | [155;157[ | [157;159[ | [159;161[ |
|---|---|---|---|---|---|
| eff. | 22 | 7 | 4 | 2 | 0 |

☞ Donner une estimation et un intervalle de confiance à 95 % du nombre moyen $\mu$ d'heures travaillées dans la population des employés.

# 2ème question

☞ Quelle est la taille minimale de l'échantillon que l'on doit choisir pour pouvoir estimer $\mu$ à 0.3 heures près (notation décimale) avec une confiance de 0.98?

# Estimation d'une variance

## Variance et écart-type corrigé

- Soit $X$ une variable quantitative prenant les valeurs $x_1, ..., x_n$ sur un échantillon de $n$ individus.
- La variance et l'écart-type de l'échantillon sont respectivement définis par :
  - $v^e = \frac{1}{n}\sum_{i=1}^{n}(x_i - m^e)^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - (m^e)^2.$
  - $s^e = \sqrt{v^e}.$
- Parfois, on voit la définition suivante de la variance ou de l'écart-type :

$$\widehat{v^e} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m^e)^2 \quad \text{et} \quad \widehat{s^e} = \sqrt{\widehat{v^e}}.$$

qui sont appelés respectivement *variance corrigée* et *écart-type corrigé*. Quelle en est la raison?

- **La variance d'un échantillon donne une estimation avec biais de la variance de la population**.
- Ceci signifie que si on prend les variances de tous les échantillons possibles et qu'on en fait la moyenne, on ne tombe pas sur la variance de la population.
- Pour corriger ce biais, on utilise *la variance corrigée*:

$$\hat{v^e} = \frac{n}{n-1} v^e = \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - m^e\right)^2.$$

# A retenir

- ☞ Si on veut utiliser un échantillon pour estimer l'écart-type d'une population : on utilise *l'écart-type corrigé* $\hat{s^e}$.
- ☞ Si on veut juste calculer l'écart-type d'un groupe : on prend le vrai écart-type $s^e$.
- ☞ Le lien entre les deux est :

$$\hat{s^e} = \sqrt{\frac{n}{n-1}} s^e$$

- ☞ Pour $n$ grand, $s^e \approx \hat{s^e}$.

# Estimation

- Soit $X$ une variable quantitative prenant les valeurs $x_1, ..., x_n$ sur un échantillon de $n$ individus issu d'une population $\mathcal{P}$.
- Le paramètre que l'on veut estimer est $\sigma = S(X)$, "écart-type" de $X$ sur la population.
- D'après ce qui précède : la meilleure estimation ponctuelle de $\sigma$ est l'écart-type corrigé $\hat{s^e}$.

- Notons $V_n$ la statistique qui donne la variance de $X$ dans les échantillons de taille $n$.
- La loi qui intervient ici est la loi du $\chi^2$.
- Introduisons la statistique

$$Y = \frac{nV_n}{\sigma^2}$$

- **Théorème d'échantillonnage de la variance** *la statistique $Y$ suit une loi du $\chi^2$ à $n-1$ ddl.*
- Rem : pour $n \leq 30$, c'est vrai uniquement si X suit une loi normale.
- C'est à partir de ce résultat qu'on construira des intervalles de confiance de l'écart-type et de la variance.

# Retour sur l'exercice 8

☞ 3) Donner un intervalle de confiance à 95 % pour l'écart-type du nombre d'heures travaillées dans ce type d'usines.