Introduction à la théorie des sondages - Cours 1

Gaspar Massiot
gaspar.massiot@ined.fr

2024-2025

Introduction

- Note
  - Devoir Maison (TD) Coefficient 1
  - Devoir sur table (CM) Coefficient 2 : Durée 1h, **Calculatrice et 1 Feuille RV autorisée**
- Intervenants
  - Cours : Gaspar Massiot (gaspar.massiot@ined.fr)
  - TD : Ulysse Lebec (ulebec@mediametrie.fr) et Nicolas Salaün (nsalaun@mediametrie.fr)
  - TP : Tony Bissonnier (tbissonnier@mediametrie.fr)
- Supports de cours inspirés de ceux de Thomas Merly-Alpa, Paul Cochet (Ined), Antoine Rebecq et Martin Chevalier (Insee)

Sommaire I

Chapitre 1

Pourquoi le sondage ?

Partie 1

Concept

Concept

Qu'est-ce que l'échantillonnage / l'estimation par sondage ?

- Une population de grande taille
- Compter ou interroger est coûteux
- On sélectionne quelques individus qui répondent ”pour tout le monde”

Idée cruciale : sélectionner **aléatoirement** ces individus.

Historique

Historiquement et conceptuellement, rien d'évident !

- Laplace (1785) : recensement par une sous-partie de la population

Historique

Historiquement et conceptuellement, rien d'évident !

- Laplace (1785) : recensement par une sous-partie de la population
- Kiaer (1895) : échantillon ”représentatif”

Historique

Historiquement et conceptuellement, rien d'évident !

- Laplace (1785) : recensement par une sous-partie de la population
- Kiaer (1895) : échantillon "représentatif"
  ... puis 1925 : acceptation de l'échantillonnage aléatoire

Historique

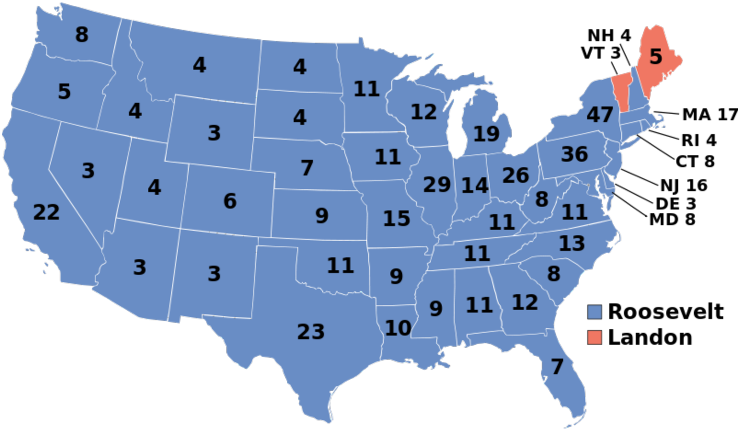
Historiquement et conceptuellement, rien d'évident !

- Laplace (1785) : recensement par une sous-partie de la population
- Kiaer (1895) : échantillon "représentatif"
  ... puis 1925 : acceptation de l'échantillonnage aléatoire
- Gallup (1936) : élections américaines

Élections américaines de 1936

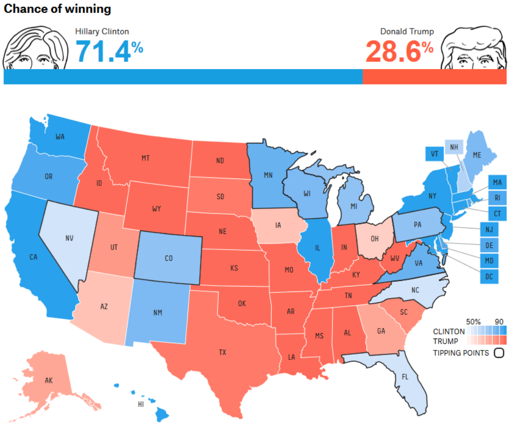
- Duel entre Alfred Landon (Républicain) et Franklin Roosevelt (Démocrate)
- Un magazine interroge ses 2 millions de lecteurices : victoire de Landon
- Gallup fait un sondage sur 50 000 personnes : il prédit la victoire de Roosevelt

Élections américaines de 1936



Élections américaines de 2016

Nate Silver, http://fivethirtyeight.com :



Élections américaines de 2016

Élection de Donald Trump, Brexit... Pourquoi les sondages ont-ils eu tout faux ?

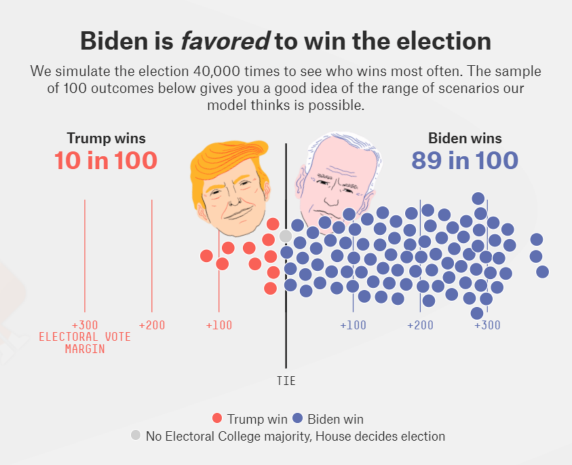
Élections américaines de 2016

Élection de Donald Trump, Brexit... Pourquoi les sondages ont-ils eu tout faux ?

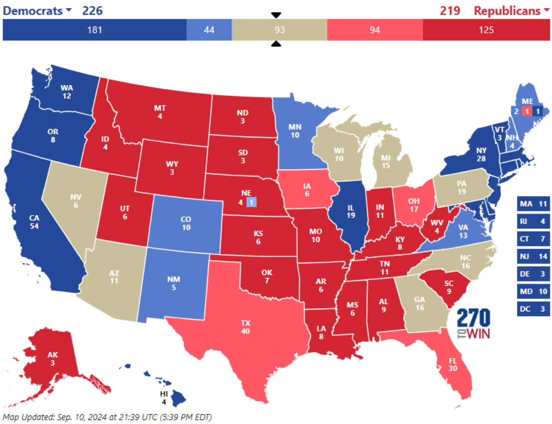
- Marge d'erreur et précision : 1 000 personnes ?
- Temporalité : changement d'avis (exemple : hausse du vote Fillon à la primaire LR 2017)
- Mensonge ou camouflage des intentions (exemple : traitement du vote FN depuis 2002)
- Ne souhaitent pas répondre : à suivre dans le cours
- Enquêtes Internet ou par téléphone : meilleurs méthodes ?

Élections américaines de 2020

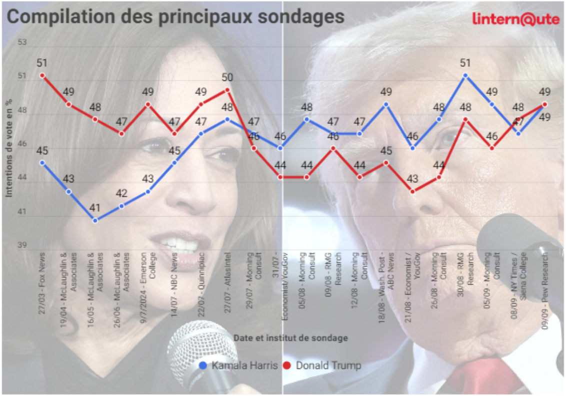
En 2020, une meilleure prédiction :



Et pour 2024 ?


Democrats 226
219 Republicans
181
44
93
94
125
WA 12
OR 8
CA 54
NV 6
ID 4
MT 4
WY 3
UT 6
AZ 11
CO 10
NM 5
ND 3
SD 3
NE 4 1
KS 6
OK 7
TX 40
MN 10
IA 6
MO 10
AR 6
LA 8
WI 10
IL 19
MI 15
IN 11
OH 17
KY 8
TN 11
MS 6
AL 9
GA 16
FL 30
SC 9
NC 16
VA 13
WV 4
PA 19
NY 28
VT 3
NH 4
ME 2 1 1
MA 11
RI 4
CT 7
NJ 14
DE 3
MD 10
DC 3
AK 3
HI 4
270 to WIN
Map Updated: Sep. 10, 2024 at 21:39 UTC (5:39 PM EDT)

Et pour 2024 ?


Compilation des principaux sondages
linternaute
Intentions de vote en %
53
51
48
46
44
41
39
27/03 - Fox News
19/04 - McLaughlin & Associates
16/05 - McLaughlin & Associates
26/06 - McLaughlin & Associates
9/7/2024 - Emerson College
14/07 - NBC News
22/07 - Quinnipiac
27/07 - AtlasIntel
29/07 - Morning Consult
31/07 - Economist/YouGov
05/08 - Morning Consult
09/08 - RMG Research
12/08 - Morning Consult
18/08 - Wash. Post - ABC News
21/08 - Economist / YouGov
26/08 - Morning Consult
30/08 - RMG Research
05/09 - Morning Consult
08/09 - NY Times / Siena College
09/09 - Pew Research
Date et institut de sondage
Kamala Harris
Donald Trump

Partie 2

Utilisations

Statistique publique

- Enquêtes auprès des ménages : le moral des ménages, le taux de chômage
- Enquêtes auprès des entreprises - ESA (Enquête Sectorielle Annuelle) : Chiffre d'affaire par secteur, chiffres d'investissement, . . .

Statistique publique

Et d'autres sujets :

- Epicov : Enquête rapide sur le covid-19 pendant les confinements (Insee-Inserm-Drees) ;
- Panel ELIPSS : Panel de sciences sociales (Sciences Po) ;
- EMP : Enquête Mobilité des Personnes (INSEE-SDES) ;
- Familles et Employeurs (Ined) . . .

Autres exemples

- Biologie : dénombrement d'espèces

Autres exemples

- Biologie : dénombrement d'espèces
- Politique

Autres exemples

- Biologie : dénombrement d'espèces
- Politique
- Marketing

Autres exemples

- Biologie : dénombrement d'espèces
- Politique
- Marketing

Pour aller plus loin sur l'utilisation des sondages et leurs limites

BLAST, Sondages d'opinion : l'Overdose

Partie 3

Pourquoi faire une enquête ?

Conception

Une enquête peut être coûteuse (en budget - 2 millions pour une enquête INSEE, mais aussi en temps des enquêté.e.s). Il faut donc s'assurer que le sujet est :

- Pertinent (contraintes européennes, demandes d'études, sujet actuel)
- Non couvert (autres enquêtes, autres données)
- Réalisable (pas trop complexe, légalité, anonymisation)

Questionnaire

Une fois les objectifs identifiés, il faut réaliser un questionnaire :

- Qui colle aux concepts
- Mais compréhensible par l'enquêté : ni équivoque, ni flou
- Qui permette de la comparabilité avec d'autres sources

$\Rightarrow$ Une étape cruciale mais difficile !

Partie 4

Un échantillon “représentatif” ?

Échantillon représentatif

Un "échantillon représentatif" : "Village" de 100 habitants

- Est-ce que le concept d'échantillon représentatif est toujours pertinent ?
- Si on veut connaître le secteur automobile en France, quelle est la bonne stratégie ?

“Sondage” devrait toujours aller de pair avec **“objectif”** (même si les objectifs pour un même échantillon peuvent être nombreux).

Secteur automobile

Quelle est le chiffre d'affaires moyen d'une entreprise du secteur automobile ?

- On a intérêt à bien interroger Renault et Peugeot.

Secteur automobile

Quelle est le chiffre d'affaires moyen d'une entreprise du secteur automobile ?

- On a intérêt à bien interroger Renault et Peugeot.
- On doit aussi interroger au hasard des garages.

Secteur automobile

Quelle est le chiffre d'affaires moyen d'une entreprise du secteur automobile ?

- On a intérêt à bien interroger Renault et Peugeot.
- On doit aussi interroger au hasard des garages.
- Ce n'est pas utile d'interroger trop de garages, car ils se ressemblent.

Secteur automobile

Quelle est le chiffre d'affaires moyen d'une entreprise du secteur automobile ?

- On a intérêt à bien interroger Renault et Peugeot.
- On doit aussi interroger au hasard des garages.
- Ce n'est pas utile d'interroger trop de garages, car ils se ressemblent.

Renault			50 Md€
Peugeot			40 Md€
Garage 1			300 k€
Garage 2			200 k€

L'estimation naïve

Pour l'estimation du total et de la moyenne d'une variable Y, l'estimateur « naïf » est :

- Pour le total, la somme des valeurs Y des individus de l'échantillon.
- Pour la moyenne, la moyenne des valeurs Y des individus de l'échantillon.

En général, l'estimation naïve est fausse (*biaisée*), surtout quand l'échantillon est choisi de façon complexe.

Secteur automobile

Quelle est le chiffre d'affaires moyen d'une entreprise du secteur automobile ?

Renault			50 Md€
Peugeot			40 Md€
Garage 1			300 k€
Garage 2			200 k€

Estimateur naïf : (50 + 40 Md + 300 + 200 k) / 4 $\approx$ 22 Md €

Partie 5

Pondération

Pondérer ?

Pour éviter d'utiliser l'estimateur naïf, on utilise généralement ce qu'on appelle des poids, qu'on note w (pour *weight* en anglais).

Le poids d'un individu correspond au nombre d'individus que l'individu de l'échantillon représente dans la population. Si l'on interroge 1 individu sur 100, le poids est alors de 100.

L'estimateur pondéré du total est alors la somme des $w_i y_i$ sur l'échantillon.

Retour sur l'exemple

Retour sur le secteur automobile. S'il n'y a qu'un Renault et qu'un Peugeot, il existe en fait près de 80 000 garages.
Les deux garages enquêtés en représentent donc 80 000 : leur poids w est de :

$$w_i = \frac{80000}{2} = 40000$$

Retour sur l'exemple

Renault	Dans l'échantillon		50 Md€
Peugeot	Dans l'échantillon		40 Md€
Garage 1	Dans l'échantillon		300 k€
Garage 2	Dans l'échantillon		200 k€
Garage 3			?
...			?
Garage 80 000			?

Retour sur l'exemple

On introduit la pondération :

Renault	Dans l'échantillon	1	50 Md€
Peugeot	Dans l'échantillon	1	40 Md€
Garage 1	Dans l'échantillon	40 000	300 k€
Garage 2	Dans l'échantillon	40 000	200 k€

Estimateur naïf : (50 + 40 Md + 300 + 200 k) / 4 $\approx$ 22 Md €

Retour sur l'exemple

On introduit la pondération :

Renault	Dans l'échantillon	1	50 Md€
Peugeot	Dans l'échantillon	1	40 Md€
Garage 1	Dans l'échantillon	40 000	300 k€
Garage 2	Dans l'échantillon	40 000	200 k€

Estimateur naïf : $(50 + 40 \text{ Md} + 300 + 200 \text{ k}) / 4 \approx 22 \text{ Md } €$
Estimateur pondéré :

$$(1*50+1*40\text{Md}+40\,000*300+40\,000*200\text{k})/(1+1+40\,000+40\,000)$$

soit environ 1,4 millions d'€

À retenir

- On construit notre sondage et donc notre échantillon dans un but précis.
- On utilise les résultats obtenus en se rappelant de notre méthode de sondage, via la pondération.

Chapitre 2

Notion d'estimateur

Partie 1

Vocabulaire

Notations - Définitions

- Population $\mathcal{U} = \{u_1, ..., u_k, ..., u_N\}$
- L'individu $u_k \in \mathcal{U}$ est repéré sans ambiguïté par son identifiant k.
- Variable d'intérêt Y, qui prend la valeur y_k pour l'individu k
- Objectif du sondage : Mesurer $\Phi(Y)$, une fonction dépendant de Y.

Notations - Définitions

Y peut être

- quantitative (exemple : revenu). Dans ce cas Φ peut être le total, la moyenne, etc.
- qualitative, c'est-à-dire prendre un nombre fini de valeurs (exemple : sexe). Dans ce cas, Φ peut être la répartition dans la population.

Notations - Définitions

La **base de sondage** donne les moyens d'identifier et de joindre les unités d'échantillonnage, souvent il s'agit des individus mais cela peut aussi être un *proxy*.

Notations - Définitions

- Échantillon $s \subset \mathcal{U}$
- Si $s = \mathcal{U}$, recensement
- Chaque individu $u_k, k \in s$ est interrogé, et on relève y_k
- Les $y_k, k \in s$ sont utilisés pour construire un **estimateur** $\hat{\Phi}$ de Φ

Plan de sondage - définition

On note $\mathcal{S}$ l'ensemble des parties de $\mathcal{U}$.
Le plan de sondage p est une loi de probabilité sur $\mathcal{S}$ i.e. :

$$\forall s \in \mathcal{S}, \ \ p(s) \geq 0$$
$$\sum_{s \in \mathcal{S}} p(s) = 1$$

Plan de sondage - exemple

Soit $\mathcal{U} = \{1, 2, 3\}$. On a alors :
$\mathcal{S} =$

Plan de sondage - exemple

Soit $\mathcal{U} = \{1, 2, 3\}$. On a alors :
$\mathcal{S} = \{\{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}$

Plan de sondage - exemple

Soit $\mathcal{U} = \{1, 2, 3\}$. On a alors :
$\mathcal{S} = \{\{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}$

On peut définir un plan de sondage p par :

$$p(\{1\}) = 0 \quad p(\{1, 2\}) = \frac{1}{2} \quad p(\{1, 2, 3\}) = 0$$

$$p(\{2\}) = 0 \quad p(\{1, 3\}) = \frac{1}{3}$$

$$p(\{3\}) = 0 \quad p(\{2, 3\}) = \frac{1}{6}$$

Paramètre d'intérêt

Y est la **variable d'intérêt** et $\Phi(Y)$ est le **paramètre d'intérêt**.

Attention, Y n'est **pas aléatoire** !

L'aléatoire repose entièrement sur l'échantillonnage décrit par le plan de sondage p.

Estimateur

Une fois l'échantillon s tiré, on **estime** $\Phi(Y)$ à l'aide d'une fonction, notée $\hat{\Phi}$, qui dépend de l'échantillon.

$\hat{\Phi}$ est appelé un **estimateur** de $\Phi(Y)$.

Espérance

$$\mathbb{E}(\hat{\Phi}) = \sum_{s} p(s) \cdot \hat{\Phi}(s)$$

C'est la valeur moyenne de $\hat{\Phi}$ obtenue avec le plan de sondage considéré **sur tous les échantillons possibles**.

Biais

$$B(\hat{\Phi}) = \mathbb{E}(\hat{\Phi}) - \Phi$$

Si $B(\hat{\Phi}) = 0$, alors on parle **d'estimateur sans biais**.

Variance / Précision

$$\mathrm{Var}(\hat{\Phi}) = \sum_{s} p(s) \cdot \left[\mathbb{E}(\hat{\Phi}) - \hat{\Phi}(s)\right]^2$$

C'est une mesure de la dispersion des valeurs $\hat{\Phi}(s)$ autour de leur moyenne.

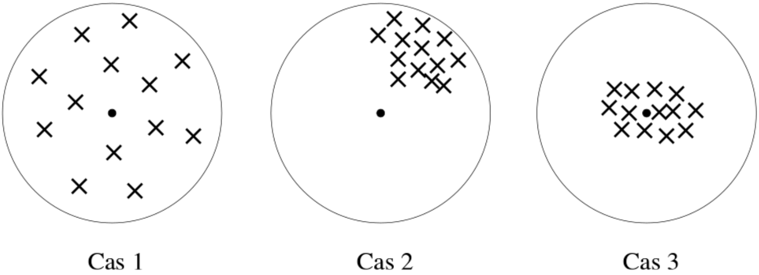
Variance / Précision

Quantités liées :

$$\sigma(\hat{\Phi}) = \sqrt{\mathrm{Var}(\hat{\Phi})}, \text{écart-type}$$

$$CV(\hat{\Phi}) = \frac{\sigma(\hat{\Phi})}{\mathbb{E}(\hat{\Phi})}, \text{coefficient de variation}$$

Schéma



Erreur quadratique moyenne

$$\begin{aligned} EQM(\hat{\Phi}) &= \sum_{s} p(s) \cdot \left[\Phi - \hat{\Phi}(s)\right]^2 \\ &= \mathrm{Var}(\hat{\Phi}) + B(\hat{\Phi})^2 \end{aligned}$$

Entre deux estimateurs sans biais, celui qui a la plus petite variance est de meilleure qualité.

Construction d'un intervalle de confiance

La **vraie variance** $\mathrm{Var}(\hat{\Phi})$ n'est pas connue (il faudrait pour cela pouvoir tirer tous les échantillons).

Il faudra donc estimer la variance à partir des données de l'échantillon. L'estimateur sera noté $\hat{V}(\hat{\Phi})$ ou $\widehat{\mathrm{Var}}(\hat{\Phi})$.

Construction d'un intervalle de confiance

Estimateurs plug-ins des quantités liées à la variance :

$$\hat{\sigma}(\hat{\Phi}) = \sqrt{\widehat{\mathrm{Var}}(\hat{\Phi})},\ \text{écart-type}$$

$$\widehat{CV}(\hat{\Phi}) = \frac{\hat{\sigma}(\hat{\Phi})}{\hat{\Phi}},\ \text{coefficient de variation}$$

Construction d'un intervalle de confiance

On fait **l'hypothèse** : $\hat{\Phi} \sim \mathcal{N}(\Phi, \mathrm{Var}(\Phi))$

L'intervalle de confiance à 95% est défini par :

$$IC_{95\%} = \left[\hat{\Phi} - 2\sigma(\hat{\Phi}); \hat{\Phi} + 2\sigma(\hat{\Phi})\right]$$

L'intervalle de confiance **estimé** est défini par :

$$\widehat{IC}_{95\%} = \left[\hat{\Phi} - 2\hat{\sigma}(\hat{\Phi}); \hat{\Phi} + 2\hat{\sigma}(\hat{\Phi})\right]$$

Partie 2

Retour sur l'estimateur naïf

L'estimateur naïf

Rappel : pour l'estimation du total et de la moyenne d'une variable Y, l'estimateur « naïf » s'écrit :

$$\hat{T}(Y)_{naif} = \sum_{k \in s} y_k$$

$$\hat{\bar{y}}_{naif} = \frac{1}{n} \sum_{k \in s} y_k$$

L'estimateur naïf

En général, l'estimation naïve est biaisée :

$$\mathbb{E}(\hat{\Phi}_{naif}) = \sum_{s} p(s) \cdot \hat{\Phi}(s)$$
$$\neq \Phi$$

$\mathbb{E}(\hat{\Phi})$ est la valeur moyenne de $\hat{\Phi}$ obtenue avec le plan de sondage considéré **sur tous les échantillons possibles**.

Partie 3

Les probabilités d'inclusion

Probabilité d'inclusion π_k

Pour améliorer l'estimateur naïf, il faut utiliser une pondération. On va calculer celle-ci à l'aide des **probabilités d'inclusion**.

La probabilité d'inclusion simple d'un individu k est la probabilité que cet individu soit dans l'échantillon. Ainsi, pour $k \in \mathcal{U}$,

$$\pi_k = \mathbb{P}(k \in s) = \mathbb{P}(\delta_k = 1) = \sum_{s \ni k} p(s)$$

où δ_k est l'indicatrice d'appartenance de k à $\mathcal{S}$, appelée aussi variable de Cornfield.

Probabilité d'inclusion π_{kl}

La probabilité d'inclusion double de deux individus k et l est la probabilité que ces deux individu soient ensemble dans l'échantillon. Ainsi, pour $k, l \in \mathcal{U}$,

$$\pi_{kl} = \mathbb{P}(k, l \in s) = \mathbb{P}(\delta_k \delta_l = 1) = \sum_{s \ni k,l} p(s)$$

Attention : on n'a pas $\pi_{kl} = \pi_k \pi_l$ en général ! On note par ailleurs $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$.

Probabilités d'inclusion π_k et π_{kl} - Propriétés

$$\mathbb{E}(\delta_k) = \pi_k \qquad \mathbb{E}(\delta_k \delta_l) = \pi_{kl}$$

$$\mathrm{Var}(\delta_\mathrm{k}) = \pi_\mathrm{k}(1 - \pi_\mathrm{k}) \qquad \mathrm{Cov}(\delta_\mathrm{k}\delta_\mathrm{l}) = \Delta_\mathrm{kl}$$

Probabilités d'inclusion π_k et π_{kl} - Propriétés

Pour un plan à **taille fixe** n, on a :

$$\sum_{k \in \mathcal{U}} \pi_k = n$$

$$\sum_{\substack{k,l \in \mathcal{U} \\ k \neq l}} \sum \pi_{kl} = n(n-1)$$

$$\sum_{\substack{l \in \mathcal{U} \\ l \neq k}} \pi_{kl} = \pi_k(n-1)$$