SAS - TD n°5

Stat. Inférentielles

Estimations ponctuelles & par intervalle, tests de conformité à un standard

Marie Ansoborlo – Ladislas Heude – Margot Periard – Emilie Tang

L’objectif de ce cours est de vous montrer comment utiliser les différentes commandes SAS pour estimer une moyenne, une fréquence, une variance et leurs intervalles de confiance ainsi que de savoir comment réaliser différents types de test de conformité à un standard. Nous illustrons ces différentes notions par des exemples basés sur le jeu de donnée “beer_recipe’.

Rappels : Un programme SAS est composé d’une ou plusieurs étapes PROC (cf TD précédents).
Ex : PROC nom_procédure data= *donnees ;*
instructions ;
RUN;

SOMMAIRE

- Estimation ponctuelle et par intervalle pour une variable qualitative
- Estimation ponctuelle et par intervalle pour une variable quantitative
- Test sur le caractère central d'une population
- Test de conformité à un standard
- Test d'indépendance entre deux variables aléatoires

1. Estimation ponctuelle et par intervalle pour une variable qualitative

PROC FREQ (1)

- s'applique aux variables qualitatives*
- arguments à préciser : DATA, TABLES…
- « Vue » des fréquences absolues, relatives et cumulées pour chaque modalité

```
PROC FREQ data=mydata;
    TABLES color_category;
run;
```

La procédure FREQ

color_category	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
pale	234	23.47	234	23.47
gold	331	33.20	565	56.67
amber	144	14.44	709	71.11
brown	79	7.92	788	79.04
black	209	20.96	997	100.00

Fréquence manquante = 3

*Prérequis SAS : ordonner les variables catégorielles (PROC FORMAT).

ex : la variable "couleur "de la bière classée en 5 catégories doit être ordonnée de "pale" à "black".

PROC FREQ (2)

- Tableau de contingence
  - croiser deux variables qualitatives pour avoir une vue sur les fréquences par sous groupe

```
PROC FREQ data=mydata;
    TABLES IBU_category * color_category;
run;
```

- Ex : fréquences des différentes catégories de “color_cat” en fonction des catégories de “IBU_category” (international bitterness unit).

Fréquence
Pourcentage
Pct de ligne
Pct de col.

Table de IBU_category par color_category

IBU_category	color_category					
	pale	gold	amber	brown	black	Total
not bitter	122 12.24 41.64 52.36	70 7.02 23.89 21.08	39 3.91 13.31 27.08	29 2.91 9.90 36.71	33 3.31 11.26 15.79	293 29.39
bit bitter	85 8.53 20.53 36.48	129 12.94 31.16 38.86	62 6.22 14.98 43.06	35 3.51 8.45 44.30	103 10.33 24.88 49.28	414 41.52
medium	18 1.81 11.69 7.73	75 7.52 48.70 22.59	24 2.41 15.58 16.67	8 0.80 5.19 10.13	29 2.91 18.83 13.88	154 15.45
quite bitter	5 0.50 8.20 2.15	22 2.21 36.07 6.63	9 0.90 14.75 6.25	5 0.50 8.20 6.33	20 2.01 32.79 9.57	61 6.12
very bitter	3 0.30 4.00 1.29	36 3.61 48.00 10.84	10 1.00 13.33 6.94	2 0.20 2.67 2.53	24 2.41 32.00 11.48	75 7.52
Total	233 23.37	332 33.30	144 14.44	79 7.92	209 20.96	997 100.00

Fréquence manquante = 3

PROC FREQ (3)

- Intervalles de confiance à 95% d'une proportion ?

  La valeur vraie du paramètre = statistique de l'échantillon ± une imprécision

  La probabilité que cette affirmation soit vraie est la confiance. L'imprécision ε est d'autant plus grande que la confiance souhaitée est élevée.

Avec : $\varepsilon \approx Z_{1-\frac{\alpha}{2}} \sqrt{\frac{P \cdot (1-P)}{n}}$

- Conditions d'applications : si nP et n(1-P) >= 5, alors la distribution de P suit la loi binomiale.

  Ex : quelle est la proportion de "very bitter" et son intervalle de confiance à 95% ? Les CA sont respectées (nP = 75 et n(1-P)= 922).

  L'estimation de la proportion de "very bitter" est de 0,07 et a 95 chances sur 100 d'être compris entre 0,06 et 0,09 soit 5 % de risque de se tromper

```
PROC FREQ data=mydata;
 tables IBU_category/binomial(level='very bitter');
run;
```

Proportion binomiale	
IBU_category = very bitter	
Proportion	0.0752
ASE	0.0084
Borne inférieure de l'IC à 95%	0.0589
Borne supérieure de l'IC à 95%	0.0916
Intervalle de confiance exact	
Borne inférieure de l'IC à 95%	0.0596
Borne supérieure de l'IC à 95%	0.0934

SAS a 2 façons différentes de calculer les intervalles de confiance, formule approchée avec le calcule de l'asymptotic standard error (ASE) soit exacte avec la loi binomiale.

2. Estimation ponctuelle et par intervalle pour une variable *quantitative*

PROC MEANS (1)

- La moyenne mais pas que…

- Statistiques descriptives simples, mots clés
  - N - effectif
  - STDDEV|STD – écart type
  - VAR - variance
  - LCLM – borne inférieure de l'IC
  - UCLM – borne supérieure de l'IC
  - NMISS - valeurs manquantes
  - CLM – Two sided Confidence Limit of the Mean
  - RANGE – Maximum minus Minimum
  - etc…

```
PROC MEANS data=mydata;
      VAR Boiltime;
run;
```

La procédure MEANS

Variable d'analyse : BoilTime

N	Moyenne	Ec-type	Minimum	Maximum
997	67.0373109	18.4763327	0	180.0000000

```
TITLE " Statistiques descriptives sur le IBU en fonction de la méthode de brassage";
PROC MEANS DATA =mydata N MEAN STDERR CLM VAR alpha=0.05;
CLASS BrewMethod;
VAR IBU;
RUN;
```

Statistiques descriptives sur le IBU en fonction de la méthode de brassage

La procédure MEANS

Variable d'analyse : IBU

BrewMethod	N obs	N	Moyenne	Erreur type	Borne inférieure de l'IC à 95% pour la moyenne	Borne supérieur de l'IC à 95% pour la moyenne	Variance
0.75	1	1	4.4800000	.	.	.	.
All Grain	698	698	45.2794126	1.7345768	41.8737907	48.6850345	2100.11
BIAB	105	105	45.3119048	3.0578733	39.2480273	51.3757822	981.8118483
N/A	1	1	27.0600000	.	.	.	.
Partial M	63	63	47.5388889	5.6440148	36.2566641	58.8211137	2006.86
extract	129	129	44.8807752	4.1728632	36.6240524	53.1374979	2246.25

- *CA : Si n >= 30 la distribution de m est approx. normale. Sinon, vérifier que la distribution de X est normale, dans ce cas le paramètre suit une distribution de Student (voir calculs)*

3.
Test sur le caractère central d'une population

Variance (σ) connue

- Pour un échantillon suffisamment grand (n>30) on suppose H0: m=m0
- **On rejette H0 si :** $|\bar{x} - m0| > \mu_{(1-\frac{\propto}{2})} \frac{\sqrt{\sigma}}{n}$

Où $\mu_{(1-\frac{\propto}{2})}$ est le quantile de la loi normale centrée réduite et $\propto$ est le risque du test

```
proc iml;
    var=4;
    qtl=quantile('NORMALE',.975);
    n=1000;
    comp=qtl*sqrt(var)/n;
    print comp;
    quit;
```

Variance (σ) inconnue

- Pour un échantillon suffisamment grand (n>30) on suppose H0: m=m0
- **On rejette H0 si :** $|\bar{x} - m0| > t_{n-1,\frac{\propto}{2}} \frac{s}{\sqrt{n}}$

Où $t_{n-1,\frac{\propto}{2}}$ est le quantile d'ordre ∝/2 de la loi de Student à n-1 degrés de liberté et $s^2 = \sum(x_i - \bar{x})^2 / (\mathrm{n} - 1)$

```
Proc ttest data=bdd_mydata h0=6.2;
    VAR ABV;
RUN;
```

La procédure TTEST

Variable : ABV

N	Moyenne	Ec-type	Err. type	Minimum	Maximum
997	6.4895	2.1749	0.0689	0.3000	27.1200

Moyenne	IC à95% - Moyenne		Ec-type	Ec.-type de l'IC à95%	
6.4895	6.3543	6.6246	2.1749	2.0835	2.2748

DDL	Valeur du test t	Pr > |t|
996	94.21	<.0001

4.
Test de conformité à un standard

Test de conformité pour une variable catégorielle

- Test du Khi² pour vérifier si les proportions observées pour une **variable catégorielle** suivent des proportions hypothétiques. Les proportions supposées sont placées entre parenthèses après l'option testp= dans la déclaration tables.
- Dans notre exemple, on veut vérifier que la proportion de All Grain est de 70% dans l'échantillon, celle de BIAB est de 10%, celle de Partial Mash de 6% et celle de extract de 14%
- H0 : les proportions sont égales
  H1 : les proportions sont différentes

```
PROC FREQ DATA = WORK.bieres;
    tables BrewMethod / chisq testp = (70 10 6 14);
RUN;
```

La procédure FREQ

BrewMethod	Fréquence	Pourcentage	Test Pourcentage	Fréquence cumulée	Pourcentage cumulé
All Grain	699	70.11	70.00	699	70.11
BIAB	105	10.53	10.00	804	80.64
Partial M	64	6.42	6.00	868	87.06
extract	129	12.94	14.00	997	100.00

Fréquence manquante = 3

Test du Khi-2 pour proportions spécifiées	
Khi-2	1.3775
DDL	3
Pr > Khi-2	0.7108

Test de normalité

Les tests de normalité permettent de vérifier si des données réelles suivent une loi normale ou non.

La procédure proc univariate suivie de l'option normal permet de vérifier si la condition de normalité est respectée.

Hypothèses statistiques :

H0 : la distribution observée est normalement distribuée

H1 : la distribution observée n'est pas normalement distribuée.

```
PROC UNIVARIATE NORMAL PLOT DATA= WORK.bieres;
    var FG;
RUN;
```

Tests de normalité				
Test	Statistique		p-value	
Kolmogorov-Smirnov	D	0.511823	Pr > D	<0.0100
Cramer-von Mises	W-Sq	5633.451	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	26100.57	Pr > A-Sq	<0.0050

Ce test est celui de Kolmogorov Smirnov, il est utilisé par le système SAS lorsque la taille de l'échantillon est relativement grande (supérieure à 2000 observations). Si la taille est plus petite, SAS utilise le test de Shapiro Wilk.

5.
Test d'indépendance entre deux variables aléatoires

Test sur le coefficient de corrélation linéaire

Le test sur le coefficient de corrélation linéaire permet de de vérifier s'il existe une dépendance linéaire significative entre deux variables quantitatives.

Hypothèses statistiques :

H0 : $p = 0$ (absence de dépendance linéaire)

H1 : $p \neq 0$ (présence d'une dépendance linéaire)

Autres hypothèses alternatives H1 :

$p > 0$ Présence d'une dépendance linéaire positive

$p < 0$ Présence d'une dépendance linéaire négative

La procédure CORR

2 Variables :	IBU ABV

Statistiques simples

Variable	N	Moyenne	Ec-type	Somme	Minimum	Maximum
IBU	73861	44.17229	42.97079	3262610	0	3409
ABV	73861	6.27194	3.17092	463252	0	156.54000

Coefficients de corrélation de Pearson, N = 73861
Proba > |r| sous H0: Rho=0

	IBU	ABV
IBU	1.00000	0.14295 <.0001
ABV	0.14295 <.0001	1.00000

```
PROC CORR DATA = WORK.bieres;
    var IBU ABV ;
RUN;
```

Le test sur le coefficient de corrélation linéaire exige que les deux populations soient normalement distribuées et que la dépendance soit linéaire. Si ces conditions ne sont pas respectées, le test sur le coefficient de corrélation des rangs de Spearman sera utilisé.

Test d'indépendance du Khi²

Un tableau croisé (tableau de contingence) permet d'analyser la relation entre deux variables qualitatives.

On peut vérifier qu'il existe une relation ou non en effectuant un test d'indépendance.

Les hypothèses statistiques sont les suivantes :

H0 : les deux variables sont indépendantes

H1 : les deux variables sont dépendantes

```
PROC FREQ DATA=WORK.bieres;
    tables SugarScale*BrewMethod / chisq;
RUN;
```

La procédure FREQ

Fréquence
Pourcentage
Pct de ligne
Pct de col.

Table de SugarScale par BrewMethod

SugarScale	BrewMethod				
	All Grain	BIAB	Partial M	extract	Total
Plato	15 1.50 100.00 2.15	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	15 1.50
Specific Gravity	684 68.61 69.85 97.85	105 10.53 10.69 100.00	64 6.42 6.52 100.00	129 12.94 13.14 100.00	982 98.50
Total	699 70.11	105 10.53	64 6.42	129 12.94	997 100.00

Fréquence manquante = 3

Statistiques pour la table de SugarScale par BrewMethod

Statistique	DDL	Valeur	Prob
Khi-2	3	6.4925	0.0900
Test du rapport de vraisemblance	3	10.7504	0.0132
Khi-2 de Mantel-Haenszel	1	5.1615	0.0231
Coefficient Phi		0.0807	
Coefficient de contingence		0.0804	
V de Cramer		0.0807	

WARNING: 38% des cellules ont un effectif théorique inférieur à 5. Le test du Khi-2 peut ne pas convenir.

Taille d'échantillon = 997
Fréquence manquante = 3

ANOVA

- Une ANOVA est utilisée lorsque l'on souhaite tester la différence des moyennes de la variable X quantitative pour chaque facteur de la variable Y qualitative.
- Dans notre jeu de données on prendra par exemple FG (Final Gravity) comme variable X et BrewMethod (méthode de brassage) comme variable Y.
- Hypothèses :
  H0 : variables indépendantes
  H1 : variables dépendantes

```
PROC ANOVA DATA = WORK.bieres;
    class BrewMethod;
    model FG = BrewMethod;
    means BrewMethod;
RUN;
```

La procédure ANOVA

Variable dépendante : FG

Source	DDL	Somme des carrés	Carré moyen	Valeur F	Pr > F
Modèle	3	0.7015136	0.2338379	1.90	0.1283
Erreur	993	122.3710224	0.1232337		
Total sommes corrigées	996	123.0725360			

R-carré	Coef de var	Racine MSE	FG Moyenne
0.005700	33.24622	0.351047	1.055899

Source	DDL	Anova SS	Carré moyen	Valeur F	Pr > F
BrewMethod	3	0.70151364	0.23383788	1.90	0.1283

Test de Kruskal-Wallis

- Le test de Kruskal-Wallis est la version non-paramétrique d'une ANOVA. On n'a pas besoin de supposer que la variable indépendante X quantitative est normalement distribuée.
- H0 : F1 = F2 = … = Fk
  H1 : il existe au moins un couple (i, j) tel que Fi ≠ Fj

```
PROC NPAR1WAY DATA = WORK.bieres;
    class BrewMethod;
    var FG;
RUN;
```

La procédure NPAR1WAY

Analyse de variance pour la variable FG
Classification par variable BrewMethod

BrewMethod	N	Moyenne
All Grain	699	1.073215
extract	129	1.015760
BIAB	105	1.013990
Partial M	64	1.016438

Source	DDL	Somme des carrés	Carré moyen	Valeur F	Pr > F
Parmi	3	0.701514	0.233838	1.8975	0.1283
Dans	993	122.371022	0.123234		

Les scores moyens ont été utilisés pour les liens.

Sitographie

https://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#statug_freq_sect028.htm

https://www.lexjansen.com/pharmasug/2003/Posters/P048.pdf

https://www.lexjansen.com/nesug/nesug08/ff/ff06.pdf

https://dms.umontreal.ca/wiki/index.php/Guide_SAS

https://stats.idre.ucla.edu/sas/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-sas/

https://stats.idre.ucla.edu/other/mult-pkg/whatstat/

MERCI !

Des questions ?