Eléments de statistique

Répétition 5

3 décembre 2019

Tests d'hypothèses

5.1. Une PME s'est spécialisée dans la production de microscopes électroniques de haute précision. Elle souhaite tester la qualité de ses produits sur base des 30 instruments produits lors d'une journée. L'ingénieur qualité décide de vérifier si la proportion de produits défectueux est supérieure à 5 % au seuil de signification α =3%.

a) Etablir la règle de décision

b) Faire de même sur base d'un échantillon de 100 microscopes électroniques.

c) Montrer que dans ce dernier cas, le test est plus puissant, en supposant que $p = 0.06$ [1].

Tests d'ajustement

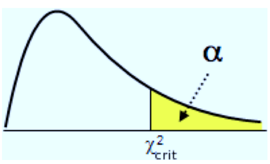
Un test d'ajustement a pour objectif de déterminer si une distribution théorique peut correspondre à une distribution de fréquences observées.

H_0 : la loi proposée convient. H_1 : la loi proposée ne convient pas.

Sous H_0, on construit une table donnant pour chaque valeur possible (ou intervalle de valeurs) de la variable étudiée l'effectif théoriquement attendu t_i, obtenu en multipliant la probabilité théorique (calculée à l'aide de la loi proposée) de cette valeur (ou intervalle de valeurs), par le nombre total n d'observations. On effectue des regroupements de lignes du tableau de manière à ce qu'aucune ligne ne comporte d'effectif théorique de taille inférieure à 5. Soit k le nombre final de lignes dans la table. On calcule ensuite la "distance" entre les effectifs observés o_i et les effectifs attendus t_i :

$$\chi^2_{obs} = \sum_{i=1}^{k} \frac{(o_i - t_i)^2}{t_i}$$

Cet estimateur suit une loi χ^2 avec un nombre de ddl (degrés de libertés) $\nu = k - 1 - a$, où a est le nombre de paramètres nécessaires à la loi qu'il a fallu obligatoirement estimer pour calculer les probabilités (1 pour une loi de Poisson, 2 pour une loi normale, ...). On rejetera finalement H_0 si la valeur de χ^2_{obs} est plus grande que la valeur critique χ^2_{crit} définie par $P(\chi^2 > \chi^2_{crit}) = \alpha$ et illustrée par la figure suivante :



La table 4 du formulaire donne les valeurs de χ^2_{crit} pour différentes valeurs typiques de α et du nombre de degrés de libertés.

1. Suggestion : calculer la puissance du test pour d'autres valeurs de p afin de pouvoir tracer une courbe comparative de l'évolution de β dans les deux cas (n=30 et n=100).

Exercices

5.2. EXAMEN AOUT 2014
On a observé le nombre $\mathcal{X}$ d'appels téléphoniques par demi-heure au central d'une entreprise.

TABLE 1: Nombre d'appels téléphoniques par tranches de 30 minutes.

Nombres d'appels	0	1	2	3	4	5	6	7	8 ou +
Observation	9	35	68	102	96	75	50	23	12

Au seuil de 5%, vérifier si ces effectifs peuvent être approchés par une loi de Poisson.

Rappel : Une variable aléatoire $\mathcal{X}$ pouvant prendre pour valeurs 0, 1, 2, ... suit une *loi de Poisson de paramètre* λ, où $\lambda \in \mathbb{R}_0^+$, si $P_{\mathcal{X}}(k) = \frac{e^{-\lambda}\lambda^k}{k!}$. Dans ce cas on a $E\{\mathcal{X}\} = V\{\mathcal{X}\} = \lambda$.

5.3. Lors de l'exercice 1.2, on avait étudié la durée de vie en heures d'un transistor d'un type bien précis sur base des résultats de la table 8. En réalisant le polygone des fréquences non cumulées, on avait constaté que ce temps semblait suivre une loi normale. Réalisez un test d'ajustement au seuil de signification de 5% afin de confirmer ou non notre impression.

TABLE 2: Durée de vie des transistors d'un fabriquant

Durée observée	Quantité
]1700,1800]	4
]1800,1900]	44
]1900,1950]	40
]1950,2000]	62
]2000,2050]	58
]2050,2100]	46
]2100,2200]	38
]2200,2300]	8

Rappels des résultats obtenus à la question 1.2 :

Moyenne estimée à l'aide des valeurs de la table :
$m_x = 0,0133 \times 1750 + 0,1467 \times 1850 + 0,1333 \times 1925 + ... = 2001,5$

Ecart-type estimé à l'aide des valeurs de la table :
$\sigma_x = \sqrt{0,0133 \times (2001,5 - 1750)^2 + 0,1467 \times (2001,5 - 1850)^2 + ...} = 102,5$

Tests d'indépendance

Un test d'indépendance a pour objectif de déterminer si deux variables qualitatives A et B, de modalités p et q respectivement, sont indépendantes.

H_0 : A et B sont indépendantes. H_1 : A et B ne sont pas indépendantes.

Sous H_0, on construit une table donnant pour chaque valeur possible de A et B, l'effectif théoriquement attendu t_{ij}, obtenu en multipliant le nombre total

n d'observations par les probabilité théorique de A_i et B_j. On effectue des regroupements de classes de manière à ce qu'aucune ligne ou colonne ne comporte d'effectif théorique de taille inférieure à 5. On calcule ensuite la distance entre les effectifs observés o_{ij} et les effectifs attendus t_{ij} :

$$\chi^2_{obs} = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{(o_{ij} - t_{ij})^2}{t_{ij}}$$

Cet estimateur suit une loi χ^2 avec un nombre de ddl $\nu = (p-1)(q-1)$. On rejetera finalement H_0 si la valeur de χ^2_{obs} est plus grande que la valeur critique χ^2_{crit} donnée par la table 4 du formulaire.

Exercice

5.4. Trois candidats briguent les suffrages des électeurs en vue d'élections législatives. Désignons les par A, B et C. Le conseiller de A décide de mener un sondage d'opinion auprès des électeurs de trois villes, que nous appelerons V_1, V_2 et V_3. Afin de pouvoir planifier au mieux la campagne électorale de A, le conseiller aimerait savoir s'ily a une différence significative entre les villes au niveau des intentions de vote. Effectuez un test d'indépendance par χ^2 au seuil de signification 0,05 sur base des résultats de la table 3.

TABLE 3: Résutats du sondage

	V_1	V_2	V_3	Total
A	50	40	35	125
B	30	45	25	100
C	20	45	20	85
Total	100	130	80	310

Exercices suggérés

5.5. Un informaticien désire se procuréer des DVD vierges à moindre coût. Un vendeur lui propose un lot de DVD dont il prétend que 5 % seulement sont défectueux. L'informaticien pense que le vendeur sous-estime la proportion de mauvais DVD. On tire un échantillon de 200 DVD dont 17 sont défectueux.

a) Au niveau de signification $\alpha = 5\%$, déterminer qui a raison.

b) Si l'acheteur pense que la proportion de DVD défectueux est en fait de 10%, déterminer le risque de l'acheteur et celui du vendeur en adoptant la règle de décision trouvée au point précédant.

c) Construire un intervalle de confiance pour le paramètre p au niveau $\alpha = 5\%$.

5.6. Une société produisant des boissons désire mettre un nouveau soda diététique sur le marché. Elle pense que son nouveau produit a meilleur goût que les sodas de la concurrence et veut vérifier cette hypothèse. Pour ce faire, elle fait goûter leur soda ainsi que 4 autres sodas à basse teneur en calories à 300 personnes. On identifie les sodas par les lettres A, B, C, D et E. Au vu des résultats de la table 4, peut-on conclure au seuil 5% qu'une des boissons a meilleur goût que les autres ?

Table 4: Nombre O_i de personne préférant la marque

Marque	A	B	C	D	E
O_i	50	65	45	70	70

5.7. EXAMEN JANVIER 2012 & AOUT 2013

On s'intéresse à la population des automobilistes et à la consommation de carburant. Un questionnaire est proposé à un échantillon de 100 conducteurs qui testeront les nouveaux pneus "Michelin Vert". La question est la suivante : "quelle est l'économie de carburant que vous constatez en l/100kms" ? On a observé une économie moyenne de carburant de 0,42 l/100kms avec un écart-type de 0,15 l/100kms.

a) Donnez un intervalle de confiance de l'économie moyenne réelle de carburant, avec un degré de confiance de 92%.

b) Selon la société Michelin, les pneus testés donnent une économie moyenne de carburant d'au moins 0,40 l/100kms. Réalisez un test unilatéral droit pour vérifiér cette affirmation au seuil $\alpha = 10\%$?

c) Quel est le risque β si, en réalité, l'économie est de 0,44 l/100kms ?

5.8. EXAMEN JANVIER 2013 & AOUT 2013

Le gérant d'un magasin de vêtements pense que le temps consacré par un vendeur à un client pendant les soldes d'hiver suit une loi exponentielle. Sur un échantillon de 200 clients, on a relevé les résultats de la table 5.

Table 5: Temps consacré par un vendeur à 200 de ses clients

Durée observée (en min)	Nombre de clients
]0,1]	68
]1,2]	45
]2,3]	29
]3,4]	15
]4,5]	15
]5,6]	7
]6,7]	9
]7,8]	6
]8,9]	3
]9,10]	2
]10,11]	1

a) Estimez le temps moyen consacré à un client.

b) Estimez les quartiles. La médiane est-elle proche de la moyenne ? Pourquoi ?

c) Au niveau de signification 5%, testez si le temps consacré à un client suit effectivement une loi exponentielle.

5.9. EXAMEN JANVIER 2014

Damien est un joueur de fléchettes débutant. Il aimerait évaluer ses capacités à lancer le plus près possible du centre de la cible. Pour ce faire, il lance 100 fléchettes et note à chaque fois à quelle distance du centre il est arrivé. On suppose que la distance ne dépasse jamais un mètre[2]. Pour plus de facilité, il regroupe ses résultats par tranche de 0,1 mètre et obtient les données de la table 6. Il montre cette dernière à son ami Stéphane, statisticien de formation.

2. La cible est accrochée au centre d'un panneau de taille 2 m x 2 m, si bien qu'il peut également mesurer sa performance lorsqu'il passe à côté de la cible.

Table 6: Résultats des 100 lancés de fléchettes

Distance par rapport au centre	Nombre de fléchettes
Entre 0 et 0,1 mètre	15
Entre 0,1 et 0,2 mètre	16
Entre 0,2 et 0,3 mètre	29
Entre 0,3 et 0,4 mètre	18
Entre 0,4 et 0,5 mètre	12
Entre 0,5 et 0,6 mètre	7
Entre 0,6 et 0,7 mètre	3
Plus de 0,7 mètre	0

a) Stéphane déclare avoir l'intuition que la distance des lancés suit une loi $\beta(2,5)$. Réalisez un test d'ajustement au seuil $\alpha = 5\%$ pour vérifier cette hypothèse.
(**15 points**)

Après mûre réflexion, Stéphane décide d'utiliser la méthode du maximum de vraisemblance pour estimer plus précisément les paramètres de la loi β. Par voie logicielle, il trouve une loi $\beta(2.136, 5.602)$. Si on effectuait un test d'ajustement sur base de cette dernière hypothèse, que peut-on affirmer sans faire le moindre calcul :

b) que la valeur χ^2_{obs} serait plus petite ou égale à celle calculée avant ? Justifiez ! (**2 points**)

c) que la valeur χ^2_{crit} serait la même qu'avant ? Sinon, que vaudrait-elle ? (**2 points**)

d) que le résultat du test serait le même qu'avant ? Justifiez ! (**1 point**)

5.10. EXAMEN JANVIER 2015
Une croyance populaire prétend qu'il y a un pic de naissances en septembre, dû au grand nombre de bébés conçus pendant les fêtes de fin d'année. En 2009, on a constaté un nombre total de 127 297 naissances en Belgique :

Mois		Nombre de naissances
Janvier	(31 jours)	10 642
Février	(28 jours)	9 638
Mars	(31 jours)	11 003
Avril	(30 jours)	10 304
Mai	(31 jours)	10 851
Juin	(30 jours)	10 423
Juillet	(31 jours)	10 927
Août	(31 jours)	10 733
Septembre	(30 jours)	10 930
Octobre	(31 jours)	10 721
Novembre	(30 jours)	10 398
Décembre	(31 jours)	10 727

a) Testez l'hypothèse "*la proportion de naissances en septembre est de* $\frac{30}{365}$" versus l'hypothèse alternative "*la proportion de naissances en septembre est plus grande que* $\frac{30}{365}$" au seuil de signification $\alpha = 5\%$. (**6 points**)

b) Quel est le risque β si en réalité la probabilité de naître en septembre est de 8,5% ? (**2 points**)

c) Si on avait travaillé avec un seuil α plus petit, le risque β aurait-il augmenté, diminué ou serait-il resté inchangé ? (**2 points**)

d) Effectuez un test d'ajustement au seuil de 5%, pour déterminer si les naissances de 2009 n'ayant pas eu lieu en septembre se répartissent de manière uniforme. Pour calculer les effectifs théoriques des différents mois, tenez bien entendu compte du nombre de jours du mois. (**10 points**)

5.11. EXAMEN AOUT 2015 & JANVIER 2016 & SEPTEMBRE 2016
Un parc d'attraction s'intéresse au nombre de visiteurs en fonction du jour de la semaine. Pour ce faire, il effectue un sondage via internet auprès de visiteurs potentiels, avec un cadeau à la clé pour les motiver. Tout d'abord, on demande aux gens s'ils sont intéressés oui ou non par une journée au parc. Si oui, on leur demande quel jour de la semaine les arrangerait le mieux. Le nombre de personnes ayant répondu oui à la première question est de 282541. Voici la préférence qu'ils ont indiquée :

Jours	Nombre de personnes préférant ce jour
Lundi	33459
Mardi	34567
Mercredi	41234
Jeudi	36789
Vendredi	42123
Samedi	47413
Dimanche	46956

a) Réalisez le diagramme en bâtonnets des fréquences observées. (**2 points**)

b) Réalisez le diagramme en escalier des fréquences cumulées du lundi au dimanche. (**2 points**)

c) Testez l'hypothèse "*la proportion de visiteurs préférant le mercredi est de* $\frac{1}{7}$" versus l'hypothèse alternative "*la proportion de visiteurs préférant le mercredi est plus grande que* $\frac{1}{7}$" au seuil de signification $\alpha = 5\%$. (**6 points**)

d) Quel est le risque β si on effectue le test décrit au point précédent, sachant que la proportion de gens préférant venir le mercredi au sein de la population complète des gens intéressés par une visite d'un jour au parc, est de 14,5% ? (**2 points**)

e) Si on avait travaillé avec un seuil α plus petit à la question (c), le risque β aurait-il augmenté, diminué ou serait-il resté inchangé ? (**2 point**)

f) On considère les personnes intéressées par une journée au parc et qui préfèrent venir le weekend. Effectuez un test d'ajustement au seuil de 5%, pour déterminer si la préférence du jour de visite se répartit de manière uniforme entre le samedi et le dimanche. (**6 points**)

5.12. EXAMEN JANVIER 2016
Une association défendant les droits des femmes s'intéresse à la probabilité qu'une personne ayant un contrat à durée déterminée (CDD) recoive un nouveau contrat à l'échéance[3]. En particulier, elle cherche à savoir s'il y a une différence entre les trois catégories suivantes : les hommes, les femmes enceintes ou ayant accouché en 2015 et les autres femmes. Elle réalise un sondage dans un grand nombre d'entreprises et obtient les résultats de la table 7.

3. Pour simplifier les choses, on dira que le CDD est prolongé dans le suite de l'énoncé. En pratique, le contrat est parfois transformé en un contrat plus intéressant pour la personne, par exemple un contrat à durée indéterminée (CDI).

Table 7: Statistiques de prolongation des CDD

	CDD prolongé	CDD non-prolongé	Total
Hommes	423	221	644
Femmes enceintes ou ayant accouché en 2015	83	64	147
Autres femmes	301	207	508
Total	807	492	1299

a) Effectuez un test d'indépendance par χ^2 au seuil de signification 0,05 pour voir s'il y a indépendance entre le fait de voir son CDD prolongé ou non et la catégorie à laquelle on appartient. (**8 points**)

Marguerite est une femme ayant un CDD. Marguerite n'est pas enceinte et n'a pas accouché en 2015.

b) Utilisez la méthode du maximum de vraisemblance pour estimer la probabilité que le CDD de Marguerite soit prolongé. (**2 points**)

c) Utilisez la méthode du maximum de vraisemblance pour construire un intervalle de confiance à 99% de la proportion de fois où le CDD de Marguerite sera prolongé. (**2 points**)

d) Utilisez l'approche bayésienne pour estimer la probabilité que le CDD de Marguerite soit prolongé. Une étude organisée par un organisme de statistique officiel sur base d'un autre échantillon suggère que la probabilité que le CDD d'une femme qui n'est pas enceinte et qui n'a pas accouché en 2015 soit prolongé puisse être modélisée a priori par une loi $\beta(19, 11)$. (**6 points**)

e) Pourquoi les méthodes utilisées aux points b) et d) ont-elles conduit quasiment au même résultat ? Qu'aurait-on pu faire pour éviter cela ? (**2 points**)

Solutions des exercices suggérés

5.5. a) On rejette H_0 si la proportion est supérieure à 7,5%, ce qui est le cas.
b) Risque de l'acheteur = 12% et du vendeur 5% c) IC = [0,0385 ; 0,1115].

5.6 $\chi^2_{obs} = 9,1667 < \chi^2_{crit} = 9,49$. On ne peut rejeter H_0.

5.7. a) [0,3936 l/100kms, 0,4464 l/100kms].
b) $0,42 > 0,4193$ on peut rejeter H_0.
c) $\beta = 0,08379$.

5.8. a) 2 minutes 27 secondes.
b) 44 secondes - 1 minute 43 secondes - 3 minutes 28 secondes - il y a beaucoup plus de petites valeurs que de grandes.
c) $\chi^2_{obs} = 5,4 < \chi^2_{crit} = 14,067$. On ne peut rejeter H_0.

5.9. a) $\chi^2_{obs} = 4,68 < \chi^2_{crit} = 11,07$. On ne peut rejeter H_0.
b) Oui, vu que les paramètres de la loi β on été optimisé pour que les effectifs théoriques s'approchent le plus possible des observations.
c) Non, on aurait du estimer deux paramètres, le nombre de ddl serait 3 et la valeur critique serait $\chi^2_{crit} = 7,81$.
d) Oui. Le test donnera un χ^2_{obs} inférieur à 4,68 et on obtiendra donc forcémment une valeur plus petite que $\chi^2_{crit} = 7,81$.

5.10. a) On peut rejeter H_0 car $0,08586 > 0,08346$.

b) 2,442%

c) Lorsque α diminue, le seuil critique augmente et β va donc augmenter.

d) $\chi^2_{obs} = 12,2 < \chi^2_{crit} = 18,3$ et on ne peut rejetter H_0.

5.11. c) On peut rejeter H_0 car 0,14594 > 0,14394.

d) 5,48%

e) Lorsque α diminue, le seuil critique augmente (il sera plus élevé que 0,14394). Le risque β va donc augmenter (il sera plus facile de passer en dessous par hasard).

f) $\chi^2_{obs} = 2,213 < \chi^2_{crit} = 3,84146$ et on ne peut rejetter H_0.

5.12. a) $\chi^2_{obs} = 7,25 > \chi^2_{crit} = 5,99$. On peut rejeter H_0.

b) 59,25%

c) [53,64% ; 64,87%]

d) 59,5%

e) Les résultats sont fort proche, car on a mis beaucoup plus de poids sur les observations que sur l'hypothèse a priori. Si on voulait augmenter le poids accordé à cette dernière, il faudrait travailler avec une autre loi, par exemple une $\beta(190, 11)$.

Rappel des solutions de l'exercice 1.2

Table 8: Table des fréquences relatives et cumulées

Durée observée	Fréquence	Fréqu. cum.
]1700,1800]	$\frac{4}{300} = 0,0133$	0,0133
]1800,1900]	$\frac{44}{300} = 0,1467$	0,16
]1900,1950]	$\frac{40}{300} = 0,1333$	0,2933
]1950,2000]	$\frac{62}{300} = 0,2067$	0,5
]2000,2050]	$\frac{58}{300} = 0,1933$	0,6933
]2050,2100]	$\frac{46}{300} = 0,1533$	0,8466
]2100,2200]	$\frac{38}{300} = 0,1267$	0,9733
]2200,2300]	$\frac{8}{300} = 0,0267$	1

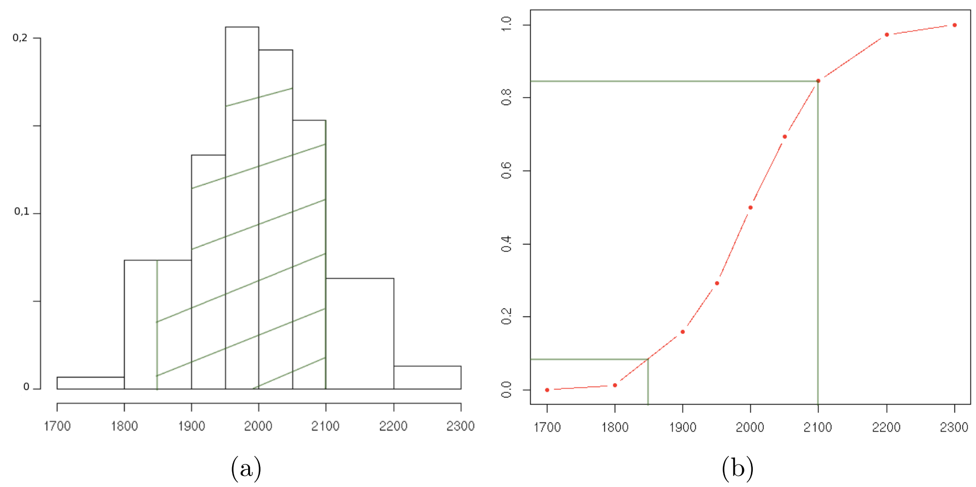


Figure 1: Histogramme (a) et polygone des fréquences cumulées (b) de la durée de vie en heures d'un transistor sur base des résultats de la table 8.