

GM – PROBABILITÉS ET STATISTIQUE

CORRIGÉS EXAMEN BLANC 1

Exercice 1 (a) L'ordre des livres n'est pas important, seulement quels livres sont lus. Il s'agit donc d'une combinaison de huit livres parmi cinq :

$$\binom{8}{5} = \frac{8!}{5! \times (8-5)!} = \frac{8 \times 7 \times 6}{3 \times 2 \times 1} = \boxed{56}$$

(b) En supposant que toutes les pièces de 5-franc sont similaires, toutes les pièces de 2-franc sont similaires, et toutes les pièces de 1-franc sont similaires, nous comptons le nombre de permutations avec éléments similaires :

$$\frac{9!}{4! \times 3! \times 2!} = \boxed{1260}$$

(c) Il y a 3 sœurs et 8 autres filles, pour un total de 11 filles. Le nombre d'arrangements de ces 11 filles d'une manière circulaire = $(11 - 1)! = 10!$. Pour voir cela, on fixe la place de la première personne. Alors, il y en a $10!$ arrangements pour les 10 personnes restantes. Puisque les gens forment un cercle, le nombre de d'arrangements possibles ne dépendent pas de la place de la première personne.

Maintenant, les trois sœurs peuvent s'arranger en $3!$ façons. Par le principe fondamental de dénombrement, le nombre de façons pour que 3 sœurs se réunissent toujours dans l'arrangement = $8! \times 3!$.

Par conséquent, le nombre de façons dont l'arrangement peut avoir lieu si les 3 sœurs ne sont PAS toutes assises ensemble est :

$$10! - (8! \times 3!) = 3,628,800 - (40,320 \times 6) = 3,628,800 - 241,920 = \boxed{3,386,880}$$

Exercice 2 (a) Soit H 'haute qualité' et W 'machine fonctionne'.

$$P(H) = P(H | W)P(W) + P(H | \text{pas } W)P(\text{pas } W) = \boxed{0.5 \times 0.9 + (1/4) \times 0.1}$$

$$(b) P(W | H) = \frac{P(H | W)P(W)}{P(H)} = \boxed{\frac{0.5 \times 0.9}{0.5 \times 0.9 + (1/4) \times 0.1}}$$

(c) Soit M 'qualité médiocre'. En conditionnant sur l'état de la machine est fonctionnant, le nombre d'objets de haute qualité suit une loi binomiale ($n = 5, p = 0.5$) [+ vérifier les 4 conditions binomiales]. Alors, selon la Formule de Bayes :

$$\begin{aligned} P(W | 4H, 1M) &= \frac{P(4H, 1M | W)P(W)}{P(4H, 1M | W)P(W) + P(4H, 1M | \text{not } W)P(\text{not } W)} \\ &= \boxed{\frac{5(1/2)^5 \times 0.9}{5(1/2)^5 \times 0.9 + 5(1/4)^4(1 - 1/4) \times 0.1}} \end{aligned}$$

Exercice 3 (a) $\int_{x=0}^1 \int_{y=0}^1 f_{X,Y}(x,y) dy dx = 1 \implies \int_{x=0}^1 \int_{y=0}^1 (x+y^2) dy dx = 1/c;$

$$1/c = \int_{x=0}^1 \left[xy + \frac{y^3}{3} \right]_{y=0}^1 dx \Rightarrow \int_{x=0}^1 (x + 1/3) dx = x^2/2 + x/3 \Big|_{x=0}^1 = 5/6 \Rightarrow \boxed{c = 6/5}$$

(b) $f_X(x) = \int_{y=0}^1 \frac{6}{5}(x+y^2) dy = \frac{6}{5} (xy + y^3/3) \Big|_{y=0}^1 = \boxed{\frac{6}{5} (x + 1/3), 0 \leq x \leq 1 \text{ (= 0 sinon)}}$

$$f_Y(y) = \int_{x=0}^1 \frac{6}{5}(x+y^2) dx = \frac{6}{5} (x^2/2 + xy^2) \Big|_{x=0}^1 = \boxed{\frac{6}{5} (1/2 + y^2), 0 \leq y \leq 1 \text{ (= 0 sinon)}}$$

(c) Non, parce que la densité conjointe ne se factorise pas ([Théorème de Factorisation](#)) :

$$\underline{f_{X,Y}(x,y) = \frac{6}{5}(x+y^2)} \quad \boxed{\neq} \quad \underline{\frac{6}{5}(x+1/3) \cdot \frac{6}{5}(1/2+y^2) = f_X(x)f_Y(y)}.$$

(d) $\underline{E[X]} = \int_{x=0}^1 x f_X(x) dx = \int_{x=0}^1 \frac{6}{5}(x^2 + x/3) dx = \frac{6}{5}(x^3/3 + x^2/6) \Big|_{x=0}^1 = \frac{6}{5} \cdot \frac{1}{2} = \underline{\underline{\frac{3}{5}}}.$

$$\underline{E[X^2]} = \int_{x=0}^1 x^2 f_X(x) dx = \int_{x=0}^1 \frac{6}{5}(x^3 + x^2/3) dx = \frac{6}{5}(x^4/4 + x^3/9) \Big|_{x=0}^1 = \frac{6}{5} \cdot \left(\frac{1}{4} + \frac{1}{9} \right) = \underline{\underline{\frac{13}{30}}}.$$

En utilisant la Formule Alternative : $\implies Var(X) = E[X^2] - (E[X])^2 = \frac{13}{30} - \left(\frac{3}{5} \right)^2 = \boxed{\underline{\underline{\frac{11}{150}}}}$

(e) $f_{X|Y}(x | y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{\frac{6}{5}(x+y^2)}{\frac{6}{5}(1/2+y^2)} = \boxed{\frac{(x+y^2)}{1/2+y^2}, 0 \leq x \leq 1, 0 \leq y \leq 1; \text{ = 0 sinon}}$

(f) $P(X < \frac{1}{2} | Y = \frac{1}{2}) = \int_{x=0}^{1/2} \frac{f_{X,Y}(x,y)}{f_Y(1/2)} dx = \int_{x=0}^{1/2} \frac{x + (\frac{1}{2})^2}{\frac{1}{2} + (\frac{1}{2})^2} dx = \frac{4}{3} \left(\frac{x^2}{2} + \frac{x}{4} \right) \Big|_{x=0}^{1/2}$
 $= \frac{4}{3} \cdot \frac{1}{4} = \boxed{\frac{1}{3}}$

Exercice 4 (a) $L(\theta) = \prod_{i=1}^n \theta x_i^{\theta-1} = \boxed{\theta^n \left(\prod_{i=1}^n X_i \right)^{\theta-1}}$

(b) $\hat{\theta}_{EMV} : \ell = n \log \theta + (\theta - 1) \sum_{i=1}^n \log X_i; \frac{d\ell(\theta)}{d\theta} = \frac{n}{\theta} + \sum_{i=1}^n \log X_i = 0 \implies \frac{n}{\theta} = - \sum_{i=1}^n \log X_i$

$$\implies \boxed{\hat{\theta}_{EMV} = - \frac{n}{\sum_{i=1}^n \log X_i}}$$

Vérifier max :

$$\frac{d^2 \ell(\theta)}{d\theta^2} = \frac{d\ell(\theta)}{d\theta} \left[\frac{n}{\theta} + \sum_{i=1}^n \log X_i \right] = -\frac{n}{\theta^2} < 0, \text{ car } n > 0, \theta > 0, \text{ donc } \frac{n}{\theta^2} > 0 \implies -\frac{n}{\theta^2} < 0,$$

alors l'extremum ($\hat{\theta}_{EMV}$) est un maximum.

$$(c) J(\theta) = \frac{-d^2 \ell(\theta)}{d\theta^2} = \boxed{\frac{n}{\theta^2}} \quad (\text{d'après la vérification de maximum, partie (b)})$$

(d) [À noter : le problème n'exige pas une évaluation du biais.]

On sait qu'asymptotiquement l'EMV est non biaisé, mais si la taille de l'échantillon est finie (petite), il faut vérifier si l'EMV est biaisé ou pas.

$$b(\hat{\theta}_{EMV}) = E[\hat{\theta}_{EMV}] - \theta; E[\hat{\theta}_{EMV}] = E \left[-\frac{n}{\sum_{i=1}^n \log X_i} \right].$$

Considérons le cas $n = 1$; si $\hat{\theta}_{EMV}$ est non biaisé, alors on a que $E \left[\frac{1}{\log X} \right] = \int_0^1 \frac{\theta x^{\theta-1}}{\log x} dx = \theta$,

et donc il nous reste à vérifier si $\int_0^1 \frac{x^{\theta-1}}{\log x} dx = 1$.

Cette intégrale est difficile à évaluer directement, mais est suffisamment compliqué qu'on peut dire que ce n'est pas le cas qu'elle égale à 1. Donc oui, $\hat{\theta}_{EMV}$ est biaisé.

(e) En supposant que $n = 9$ est 'suffisamment grande', un IC approximatif à 95% pour θ est (ok si 2 au lieu de 1.96) : $\hat{\theta}_{EMV} \pm 1.96/\sqrt{J(\hat{\theta}_{EMV})}$

$$\implies \hat{\theta}_{EMV} \pm 1.96/\sqrt{\hat{\theta}^2/n} \implies -\frac{n}{\sum_{i=1}^n \log X_i} \pm 1.96 \times \sqrt{n}/\left| \sum_{i=1}^n \log X_i \right|.$$

$$\sum_{i=1}^n \log X_i = \log e^{-1} + \log e^{-2} + \log e^{-4} + \log e^{-2} + \log e^{-1} + \log e^{-3} + \log e^{-3} + \log e^{-1} + \log e^{-1}$$

$$= (-1) + (-2) + (-4) + (-2) + (-1) + (-3) + (-3) + (-1) + (-1) = -18;$$

$$\text{alors l'IC est } 9/18 \pm 2 \times \sqrt{9}/18 \implies \boxed{1/2 \pm 1/3}$$

Exercice 5 A : ANOVA

(a) Le nombre de D groupes est $df+1 = 1+1 = 2$, et le nombre de W groupes est $df+1 = 2 + 1 = 3$, donc le tableau est 2×3 . Dans ce cas le nombre de df interaction ($D \times W$) est $(2-1) \times (3-1) = 1 \times 2 = \mathbf{2}$.

(b) Total $df = 1 + 2 + 2 + 18 = \mathbf{23}$ (donc le nombre total d'observations = $23 + 1 = 24$).

(c) SC Erreur = CM Erreur $\times$ Erreur $df = 2 \times 18 = 36$ (ou : SC Erreur = $272 - 20 - 200 - 16 = 36$).

(d) CM $D = \text{SC } D / df D = 20/1 = 20$.

(e) CM $D \times W = \text{SC } D \times W / df D \times W = 16/2 = 8$.

(f) $F D \times W = \text{CM } D \times W / \text{CM Erreur} = 8/2 = 4$.

(g) De la partie **(b)**, le nombre total d'observations est 24, et le nombre de combinaisons $D \times W$ est $2 \times 3 = 6$. Donc, le nombre total d'observations par combinaisons est $24/6 = 4$.

(h) Les lignes NE SONT PAS PARALLÈLES, qui suggère la présence d'une interaction.

(i) 1. paramètres : les paramètres d'interaction, $\gamma_{i,j}, i = 1, 2; j = 1, 2$.

2. $H : \gamma_{i,j} = 0, \forall i, j$

$A : \text{Au moins un } \gamma_{i,j} \neq 0, i = 1, 2; j = 1, 2$

3. Statistique de test F_{obs} pour $D \times W = 4$ (partie **f**)

4. Sous $H, F \sim F_{2,18} \implies p_{obs} = 0.037$ (d'après le tableau)

5. Pour $p_{obs} = 0.037 < 0.05 = \alpha \implies$ REJETTE H , il y a une *interaction statistiquement significative* entre détergent D et température de l'eau W .

La conclusion est que l'effet combiné du détergent et la température de l'eau n'est pas la même que la somme des effets du détergent (individuel) et l'effet de la température de l'eau. Vous ne pouvez pas vous attendre à la même efficacité détergente quel que soit la température de l'eau (ou, de manière équivalente, la même efficacité de température de l'eau pour chacun des détergents).

B : RÉGRESSION

(a) Afin de tester si le modèle en gros est utile, il nous faut un F -test global.

1. paramètres : les $\beta_i, i = 1, \dots, 9$ dans la population

2. $H : \beta_1, \dots, \beta_9 = 0$

[Aucune des variables n'aide pas pour expliquer la fréquentation aux parcs StatisticsLand.]

$A : \text{Au moins un } \beta_i, i = 1, \dots, 9 \neq 0$

[Au moins une des variables est utile pour expliquer la fréquentation ; le modèle en gros fait un bon travail pour l'explication du nombre de visiteurs aux parcs StatisticsLand.]

3. $F_{obs} = 265.5$ (du tableau d'ANOVA correspondant).

4. Sous $H, F \sim F_{9,110}; p_{obs} \approx 0$ (du tableau d'ANOVA correspondant).

5. Puisque $p_{obs} = 0 < 0.05 = \alpha$, alors on REJETTE l'hypothèse NULLE et en conclut qu'au moins une des 9 variables aide en l'explication du nombre de visiteurs aux parcs StatisticsLand.

(b) Pour décider si un modèle fait de *bonnes prédictions* nous devons regarder l'erreur 'typique' (la REQM) et la comparer aux valeurs Y que nous voulons prédire. Ici, le REQM = 0,2, ce qui signifie que la prédiction est généralement en erreur par environ 200 000 personnes du nombre de visiteurs d'un parc StatisticsLand dans un trimestre donné. Puisque le nombre moyen de visiteurs $\bar{Y} = 5$ (soit 5 millions de visiteurs), le modèle semble fournir une assez bonne prédiction (en erreur par environ 200 000 seulement de 5 millions).

(c) Le coefficient de la variable Rides (manèges) est $b_3 = 0,01$. Cela signifie que pour chaque trajet supplémentaire au parc, la fréquentation devrait augmenter de 0,01 million = 10 000 personnes, en supposant que toutes les autres variables sont maintenues fixes.

Le coefficient de la variable Discount (réduction) est $b_7 = 1$. Cela nous indique qu'en moyenne, lorsqu'il y a une remise étant offert 1 million de personnes supplémentaires viennent au parc par trimestre que lorsqu'il n'y a pas de rabais, en supposant que toutes les autres variables sont maintenues fixes.

(d) La variable Size X_4 a une corrélation assez forte avec la fréquentation, Y , il devrait donc en soi être un bon prédicteur. Comme la corrélation est positive, on s'attendrait à b_4 doit également être positif, ce qui indique que la Size augmente ainsi que la fréquentation. Cela a du sens dans le monde réel : plus le parc est grand, plus il peut contenir de personnes.

(e) On simplement fait la substitution des valeurs en l'équation de la régression (attention aux unités). Les valeurs ont été mesurées 4 fois par an pendant 5 année, alors il y a 20 points avant cet an. Été (Summer) est le 3ème quart de l'année (hiver, printemps, été, automne) donc ce temp est le numéro 23. La population est mesurée en centaines de milliers de gens, donc on a $X_5 = 50$ pour 5 million (gens). Il n'y a pas de rabais, alors $X_7 = 0$ et le parc se trouve en romandie qui nous dit que $X_8 = X_9 = 0$. Les valeurs restantes peuvent être substituées comme elles sont données, qui nous donne :

$$\hat{Y} = 2 + 0.25(23) - 0.2(45) + 0.01(50) - 0.001(50) + 0.05(50) + 0.1(70) + 0 - 2(0) - 1(0) = 8.7.$$

Donc, la prédiction est que le nombre de visiteurs serait 8.7 million cet été.

(f) Le coefficient du Price (prix) est de -0,2, ce qui signifie que pour chaque chf supplémentaire facturé on prévoit une baisse de fréquentation de 200 000 personnes, toutes choses égales par ailleurs. Ainsi, une augmentation de 5 chf correspondrait à une perte prévue de 1 million de personnes.

(g) 1. paramètre - β_6 , le coefficient pour température dans la population

2. $H : \beta_6 = 0$: Une fois les autres variables sont prises en compte, la température n'a pas une association (ou même une relation négative) avec la fréquentation.

$A : \beta_6 > 0$: Une fois les autres variables sont prises en compte, la température a une association / relation positive avec la fréquentation.

3. $t_{obs} = 2.08$ (la t -statistique pour température).

4. Sous H , $t \sim t_{110} \implies p_{obs} = 0.0396/2 = 0.0198$.

[Puisque le test est unilatéral, il faut diviser la p -valeur des sorties en deux, qui nous donne 0.0198.

5. $p_{obs} = 0.0198 < 0.05 = \alpha$, on REJETTE l'hypothèse NULLE H , et *en conclut qu'une température plus élevée est associée à une plus grande fréquentation, même après avoir pris en compte toutes les autres variables*. On conseille donc au professeur Sadisticus d'ajouter le nouveau manège et les stands vendant la glace.

GM – PROBABILITÉS ET STATISTIQUE

PRACTICE FINAL EXAM 1 SOLUTIONS

Exercise 1 (a) The order of the books is not important, only which books are read. This is a combination of eight items taking five at a time :

$$\binom{8}{5} = \frac{8!}{5! \times (8-5)!} = \frac{8 \times 7 \times 6}{3 \times 2 \times 1} = \boxed{56}$$

(b) Assuming that all 5-franc pieces are similar, all 2-franc pieces are similar, and all 1-franc pieces are similar, we are counting the number of permutations with similar elements :

$$\frac{9!}{4! \times 3! \times 2!} = \boxed{1260}$$

(c) There are 3 sisters and 8 other girls for a total of 11 girls. The number of ways to arrange these 11 girls in a circular manner = $(11-1)! = 10!$. To see this, fix the place of the first person. Then there are $10!$ orders for the remaining 10 people. Since the people are in a circle, the number of possible orders does not depend on the place of the first person.

Now, the three sisters can rearrange themselves in $3!$ ways. By the basic counting principle, the number of ways so that 3 sisters always come together in the arrangement = $8! \times 3!$.

Therefore, the number of ways in which the arrangement can take place if the 3 sisters are NOT all seated together is :

$$10! - (8! \times 3!) = 3,628,800 - (40,320 \times 6) = 3,628,800 - 241,920 = \boxed{3,386,880}$$

Exercise 2 (a) Let H denote ‘high quality’ and W denote ‘machine works’.

$$P(H) = P(H | W)P(W) + P(H | \text{not } W)P(\text{not } W) = \boxed{0.5 \times 0.9 + (1/4) \times 0.1}$$

$$(b) P(W | H) = \frac{P(H | W)P(W)}{P(H)} = \boxed{\frac{0.5 \times 0.9}{0.5 \times 0.9 + (1/4) \times 0.1}}$$

(c) Let M denote ‘medium quality’. Conditional on the machine working, the number of high quality items has a binomial ($n = 5, p = 0.5$) distribution [+ verify the 4 binomial conditions].

Then, by Bayes’ Rule :

$$\begin{aligned} P(W | 4H, 1M) &= \frac{P(4H, 1M | W)P(W)}{P(4H, 1M | W)P(W) + P(4H, 1M | \text{not } W)P(\text{not } W)} \\ &= \boxed{\frac{5(1/2)^5 \times 0.9}{5(1/2)^5 \times 0.9 + 5(1/4)^4(1 - 1/4) \times 0.1}} \end{aligned}$$

Exercise 3 (a) $\int_{x=0}^1 \int_{y=0}^1 f_{X,Y}(x, y) dy dx = 1 \implies \int_{x=0}^1 \int_{y=0}^1 (x + y^2) dy dx = 1/c;$

$$1/c = \int_{x=0}^1 \left[xy + \frac{y^3}{3} \right]_{y=0}^1 dx \implies \int_{x=0}^1 (x + 1/3) dx = x^2/2 + x/3 \Big|_{x=0}^1 = 5/6 \implies \boxed{c = 6/5}$$

$$(b) f_X(x) = \int_{y=0}^1 \frac{6}{5}(x+y^2) dy = \frac{6}{5} (xy + y^3/3) \Big|_{y=0}^1 = \boxed{\frac{6}{5} (x + 1/3), 0 \leq x \leq 1 \quad (= 0 \text{ otherwise})}$$

$$f_Y(y) = \int_{x=0}^1 \frac{6}{5}(x+y^2) dx = \frac{6}{5} (x^2/2 + xy^2) \Big|_{x=0}^1 = \boxed{\frac{6}{5} (1/2 + y^2), 0 \leq y \leq 1 \quad (= 0 \text{ otherwise})}$$

(c) No, because the joint density does not factor (**Factorization Theorem**) :

$$\underline{f_{X,Y}(x, y) = \frac{6}{5}(x + y^2)} \quad \boxed{\neq} \quad \underline{\frac{6}{5}(x + 1/3) \cdot \frac{6}{5}(1/2 + y^2) = f_X(x)f_Y(y)}.$$

$$(d) \underline{E[X]} = \int_{x=0}^1 x f_X(x) dx = \int_{x=0}^1 \frac{6}{5}(x^2 + x/3) dx = \frac{6}{5}(x^3/3 + x^2/6) \Big|_{x=0}^1 = \frac{6}{5} \cdot \frac{1}{2} = \underline{\frac{3}{5}}.$$

$$\underline{E[X^2]} = \int_{x=0}^1 x^2 f_X(x) dx = \int_{x=0}^1 \frac{6}{5}(x^3 + x^2/3) dx = \frac{6}{5}(x^4/4 + x^3/9) \Big|_{x=0}^1 = \frac{6}{5} \cdot \left(\frac{1}{4} + \frac{1}{9}\right) = \underline{\frac{13}{30}}.$$

$$\text{Using the Alternative Formula : } \Rightarrow \text{Var}(X) = E[X^2] - (E[X])^2 = \frac{13}{30} - \left(\frac{3}{5}\right)^2 = \boxed{\frac{11}{150}}$$

$$(e) f_{X|Y}(x | y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} = \frac{\frac{6}{5}(x + y^2)}{\frac{6}{5}(1/2 + y^2)} = \boxed{\frac{(x + y^2)}{1/2 + y^2}, 0 \leq x \leq 1, 0 \leq y \leq 1; \quad = 0 \text{ otherwise}}$$

$$(f) P(X < \frac{1}{2} | Y = \frac{1}{2}) = \int_{x=0}^{1/2} \frac{f_{X,Y}(x, y)}{f_Y(1/2)} dx = \int_{x=0}^{1/2} \frac{x + (\frac{1}{2})^2}{\frac{1}{2} + (\frac{1}{2})^2} dx = \frac{4}{3} \left(\frac{x^2}{2} + \frac{x}{4} \right) \Big|_{x=0}^{1/2} \\ = \frac{4}{3} \cdot \frac{1}{4} = \boxed{\frac{1}{3}}$$

$$\textbf{Exercise 4} \quad (a) L(\theta) = \prod_{i=1}^n \theta x_i^{\theta-1} = \boxed{\theta^n \left(\prod_{i=1}^n X_i \right)^{\theta-1}}$$

$$(b) \hat{\theta}_{EMV} : \ell = n \log \theta \times (\theta - 1) \sum_{i=1}^n \log X_i ; \frac{d\ell(\theta)}{d\theta} = \frac{n}{\theta} + \sum_{i=1}^n \log X_i = 0 \Rightarrow \frac{n}{\theta} = - \sum_{i=1}^n \log X_i \\ \Rightarrow \boxed{\hat{\theta}_{EMV} = - \frac{n}{\sum_{i=1}^n \log X_i}}$$

Verify max :

$$\frac{d^2 \ell(\theta)}{d\theta^2} = \frac{d\ell(\theta)}{d\theta} \left[\frac{n}{\theta} + \sum_{i=1}^n \log X_i \right] = -\frac{n}{\theta^2} < 0, \text{ car } n > 0, \theta > 0, \text{ thus } \frac{n}{\theta^2} > 0 \Rightarrow -\frac{n}{\theta^2} < 0,$$

so the extremum ($\hat{\theta}_{MLE}$) is a maximum.

$$(c) J(\theta) = \frac{-d^2 \ell(\theta)}{d\theta^2} = \boxed{\frac{n}{\theta^2}} \quad (\text{according to the verification of the maximum, part (b)}).$$

(d) [Note : the problem does not ask us to evaluate the bias.]

We know that asymptotically the MLE is unbiased, but if the sample size is finite (small) we must verify whether the MLE is biased or not.

$b(\hat{\theta}_{MLE}) = E[\hat{\theta}_{MLE}] - \theta$; $E[\hat{\theta}_{MLE}] = E\left[-\frac{n}{\sum_{i=1}^n \log X_i}\right]$. Consider the case $n = 1$; if $\hat{\theta}_{MLE}$ is unbiased, then we have $E\left[\frac{1}{\log X}\right] = \int_0^1 \frac{\theta x^{\theta-1}}{\log x} dx = \theta$, and thus it remains to verify whether $\int_0^1 \frac{x^{\theta-1}}{\log x} dx = 1$. This integral is difficult to evaluate directly, but is sufficiently complicated that we can say that it is not the case that it is equal to 1. Thus yes, $\hat{\theta}_{EMV}$ is biased.

(e) Assuming that $n = 9$ is ‘sufficiently large’, an approximate 95% CI for θ is (ok to use 2 instead of 1.96) : $\hat{\theta}_{MLE} \pm 1.96/\sqrt{J(\hat{\theta}_{MLE})}$

$$\Rightarrow \hat{\theta}_{MLE} \pm 1.96/\sqrt{\hat{\theta}^2/n} \Rightarrow -\frac{n}{\sum_{i=1}^n \log X_i} \pm 1.96 \times \sqrt{n}/\left|\sum_{i=1}^n \log X_i\right|$$

$$\begin{aligned} \sum_{i=1}^n \log X_i &= \log e^{-1} + \log e^{-2} + \log e^{-4} + \log e^{-2} + \log e^{-1} + \log e^{-3} + \log e^{-3} + \log e^{-1} + \log e^{-1} \\ &= (-1) + (-2) + (-4) + (-2) + (-1) + (-3) + (-3) + (-1) + (-1) = -18; \end{aligned}$$

$$\text{so the CI is } 9/18 \pm 2 \times \sqrt{9}/18 \Rightarrow \boxed{1/2 \pm 1/3}$$

Exercise 5 A : ANOVA

(a) The number of D groups is $df+1 = 1+1 = 2$, and the number of W groups is $df+1 = 2 + 1 = 3$, so the table is 2×3 . In this case the number of df interaction ($D \times W$) is $(2 - 1) \times (3 - 1) = 1 \times 2 = \mathbf{2}$.

(b) Total $df = 1 + 2 + 2 + 18 = \mathbf{23}$ (so total number of observations = $23 + 1 = 24$).

(c) Error SS = Error MS $\times$ Error $df = 2 \times 18 = 36$ (or : Error SS = $272 - 20 - 200 - 16 = 36$).

(d) MS $D = \text{SS } D / df \text{ } D = 20/1 = 20$.

(e) MS $D \times W = \text{SS } D \times W / df \text{ } D \times W = 16/2 = 8$.

(f) $F \text{ } D \times W = \text{MS } D \times W / \text{MS Error} = 8/2 = 4$.

(g) From part (b), the total number of observations is 24, and the number of combinations $D \times W$ is $2 \times 3 = 6$. Thus, the number of observations per combination is $24/6 = 4$.

(h) The lines are NOT PARALLEL, suggesting the presence of interaction

(i) 1. parameters : the interaction parameters, $\gamma_{i,j}, i = 1, 2; j = 1, 2$.

2. $H : \gamma_{i,j} = 0, \forall i, j$

$A : \text{At least one } \gamma_{i,j} \neq 0, i = 1, 2; j = 1, 2$

3. Test statistic F_{obs} for $D \times W = 4$ (part f)

4. Under $H, F \sim F_{2,18} \Rightarrow p_{obs} = 0.037$ (from table)

5. For $p_{obs} = 0.037 < 0.05 = \alpha \Rightarrow \text{REJECT } H$, there is a *statistically significant interaction* between detergent D and water temperature W .

The conclusion is that the combined effect of detergent and water temperature is not the same as the sum of the individual detergent effect and water temperature effect. You cannot expect the same detergent efficacy regardless of water temperature (or, equivalently, the same efficacy of water temperature for each of the detergents).

B : Regression

(a) To test whether the model as a whole is useful we need an **overall F -test / F -test global**.

1. parameters : the $\beta_i, i = 1, \dots, 9$ in the population

2. $H : \beta_1, \dots, \beta_9 = 0$

[None of the variables is helpful for explaining the attendance at StatisticsLand parks.]

$A : \text{At least one } \beta_i, i = 1, \dots, 9 \neq 0$

[At least one of the variables is useful for explaining attendance ; the model as a whole does a good job of explaining the number of visitors at the theme parks].

3. $F_{obs} = 265.5$ (from the corresponding ANOVA table).

4. Under H , $F \sim F_{9,110}; p_{obs} \approx 0$ (from the corresponding ANOVA table).

5. Since $p_{obs} = 0 < 0.05 = \alpha$, we REJECT the null hypothesis and conclude that at least one of the nine variables helps explain the attendance at the parks.

(b) To decide whether a model makes *good predictions* we must look at the ‘typical’ error (the RMSE) and compare it to the Y values we are trying to predict. Here, the $\text{RMSE} = 0.2$, meaning that we typically are off by about 200,000 people when we try to predict the number of visitors to a StatisticsLand park in a given quarter. Since the average number of visitors $\bar{Y} = 5$ (that is, 5 million visitors), the model seems to provide a pretty good prediction (off by around only 200,000 from 5 million).

(c) The coefficient for the Rides variable is $b_3 = 0.01$. This means for every additional ride there is at the park, attendance is expected to go up by 0.01 million = 10,000 people, assuming all the other variables are held fixed.

The coefficient for the discount variable is $b_7 = 1$. This tells us that on average, when there is a discount being offered 1 million more people come to the park per quarter than when there is not a discount, assuming all other variables are held fixed.

(d) The size variable X_4 has a quite strong correlation with attendance, Y , so it should by itself be a good predictor. Since the correlation is positive we would expect b_4 to be positive as well, indicating that as size increases so does attendance. This makes real-world sense : the bigger the park, the more people it can hold.

(e) We just plug into the regression equation (being careful of the units). The data have been measured quarterly for 5 years so there have been 20 time points before this year. Summer is the 3rd quarter of the year (winter, spring, summer, fall) so this must be time point 23. The population is measured in hundreds of thousands of people so we have $X_5 = 50$ for 5 million people. There is no discount offered so $X_7 = 0$ and the park is in romandie so $X_8 = X_9 = 0$. The rest of the numbers can be plugged in as they are given, yielding

$$\hat{Y} = 2 + 0.25(23) - 0.2(45) + 0.01(50) - 0.001(50) + 0.05(50) + 0.1(70) + 0 - 2(0) - 1(0) = 8.7.$$

Thus we predict the park will have 8.7 million customers this summer.

(f) The coefficient for the price is -0.2, meaning that for every extra chf charged there is an expected decrease in attendance of 200,000 people, all else being equal. Thus a 5 chf raise would correspond to an expected loss of 1 million people.

- (g) 1. parameter - β_6 , the coefficient for temperature in the population
2. $H : \beta_6 = 0$: Once the other variables have been taken into account, temperature has no association (or even a negative relationship) with attendance.
- $A : \beta_6 > 0$: After accounting for the other factors, temperature has a positive association / relationship with attendance.
3. $t_{obs} = 2.08$ (the t -statistic for temperature).
4. Under H , $t \sim t_{110} \implies p_{obs} = 0.0396/2 = 0.198$.
[Since the test is 1-sided we have to divide the p-value from the printout in half yielding 0.0198.]
5. $p_{obs} = 0.0198 < 0.05 = \alpha$, we REJECT the null hypothesis H , and *conclude that higher temperatures are associated with higher attendance, even after accounting for all the other variables*. We would therefore advise Professor Sadisticus that he should add the new ride and ice cream stands.