22 Aout   
Il y a deux parties aux stats:  
- stats descriptive: utiliser les données disponibles pour décrire la population et/ou échantillon en autant que possible  
- inférence statistique: quelles conclusions peut être tirés des stats descriptives, et avec quel niveau de confiance  
  
Termes stats :  
- Une population est l’ensemble de tous les objets nous intéressant. Un individu est un membre de cette population. Un échantillon est le sous-ensemble que l’on étudie directement.   
Exemple : Je prends votre taille pour extrapoler à l’ensemble des étudiants du CVM. (Population = tous les étudiants du cégep, Échantillon = notre classe)  
Exemple : On test un programme sur cent listes de données pour tester son efficacité. (Échantillon = ensemble des listes testées, Population = ensemble de toutes les données possibles qui pourraient hypothétiquement être utilisées, Variable statistique = le temps que le programme prendrait pour traiter chaque liste de données)  
  
- Une variable statistique est une propriété que tout individu de la population a mais dont la valeur varie d’un individu à l’autre.  
  
- Une variable qualitative : non-numérique   
  
- Une variable quantitative discrète : si les valeurs qu’elle peut prendre ne forment pas des intervalles sur l’axe des réels (Remarque : 95% du temps les variables discrètes sont numériques et ne peuvent prendre que des valeurs entières)  
  
- Une variable quantitative continue : numérique, peut prendre n’importe quelle valeur  
  
Exemple : 20 essais d’un programme, on note le temps qu’il prend à accomplir sa tache et s’il y a eu des problèmes de performances.  
Variables : Temps (variable quantitative continue) et problèmes (variable qualitative)  
Dépouillement :  
- Problèmes : Aucun 18, mineur 2, fatal 0  
-  
  
Exemple : 30 étudiants dans une classe, on note la taille et l’âge.  
Variables : Taille (variable quantitative continue) et âge (variable quantitative discrète, mais peut techniquement être continue si on y va très précisément)  
Dépouillement :  
-   
-  
  
  
  
Premièrement avec ces données (Dépouillement des données) :  
- Qualitatives : compter le nombre d’individu pour chaque valeur possible de la variable  
- Quantitatives continues : diviser en un nombre d’intervalles et compter le nombre d’individus pour chaque intervalle. (Pour la présentation des données) # de classes à peu près = 1+3.322logn où n est la taille de l’échantillon… en pratique, chercher des limites significatives et rondes.   
- Quantitatives discrètes : l’une ou l’autre des solutions présenter précédemment selon les données que l’on a. Dépendamment s’il y a beaucoup de valeur possible ou pas.  
  
Définition : Si on a un intervalle de données [a,b[, a et b en sont les limites, (a+b)/2 le point milieu. La fréquence absolue d’un intervalle et/ou valeurs est le nombre de fois qu’elle apparait, fréquence relative : la proportion.  
  
Présentation préféré :  
1. Variable qualitative  
- Fréquence absolues : Diagramme en bâtons (pas vraiment d’autres options)  
- Fréquence relative : Diagramme de secteurs (idéal pour représenter des proportions)  
2. Variable quantitative   
- Fréquence relative : Diagramme de secteurs  
- Fréquence absolue : Histogramme (important de bien annoté l’axe des x grâce à la valeur de la variable), diagramme à bâton si valeurs exactes (discrète) ou polygone de fréquence encore la si valeur exacte (discrète)  
(Le but de ces graphiques est d’approximé des fonctions)

29 Aout   
Mesures de tendance centrale  
  
Définition 1 : Un mode d’une variable statistique est la ou une des valeurs a la plus grande fréquence absolue. (Remarque : On peut avoir plus qu’un mode, mais uniquement si deux valeurs on exactement la même fréquence absolue.)   
Définition 2 : Si X est une variable quantitative, la moyenne (μx ou ¯x¯, voir notes de cours papier pour plus de clarté) est la valeur que la variable prendrait si elle était constante sur notre échantillon mais la somme de la valeur serait la même. (Remarque : Très affectée par les valeurs extrême)  
Définition 3 : La médiane, notée Md(x), est la valeur approximative pour laquelle la moitié de l’échantillon se trouve au-dessus et l’autre moitié en-dessous. (Remarque : 1. Le calcul de la médiane est plus difficile 2. Parfois mal défini)  
  
Cas 1 : X est continue ou discrète avec beaucoup de modalité.   
- Le mode ne nous intéresse pas dans ce cas.  
- Calcul de la moyenne ; Supposons comme suit : N individus pour lesquels X prend les valeurs x1 ,x2 , …, xN , avec répétitions possibles. Voir NdeC p.2 pour formules complète.  
- Calcul de la médiane ; On suppose que x1 , x2 , …, xN , sont en ordre croissant où N est pair :   
Md(x)=(xn /2 +xn/2+1 )/2, où N est impair : Md(x)=x(N+1)/2

Moy>Méd : valeurs aberrantes très grande  
Méd>Moy : valeurs aberrantes très petite

12 septembre  
  
2ième situation : X est une variable quantitative discrète pouvant prendre un nombre restreint de valeur.  
  
Les variables : Population à l’échantillon de taille N. X peut prendre la valeurs x1, x2, …, xk. La fréquence absolue n1, n2, …, nk. Ni = # d’individus pour lesquels x=Xi. La fréquence relative f1, f2, …, fk, fi=ni/N. Finalement, la fréquence relative cumulée, F1, F2, …, Fk, Fi = f1+f2+…+fk.

S’il n’y a pas de valeur pour lequel Fi = 0.5 pour pouvoir calculer la médiane, voir notes p.3

3ième situation : Variable avec valeurs regroupés en classes

X variable, N individu, regroupés en classe.  
[b1,b2[, [b2,b3[, …, [bk,bk+1[  
mi, point milieu de la i-ème classe, pour mi=(bi+bi+1)/2  
On va supposer que la moyenne à l’intérieur d’une classe est le mi  
Pour médiane voir notes de cours p.5, legit c’est pas si pire. Sauf si une classe a un Fi=0.5, prendre la valeur supérieure de l’intervalle sans calcul.  
  
Mesure de dispersion

Une mesure de dispersion tente de mesurer à quel point une variable varie.  
Définition 1: Si X est une variable statistique, l’étendue est la différence entre la plus grande valeur de X et de la plus petite.  
Définition 2 : Si X est une variable numérique sur une **population** avec une moyenne ¯x¯, sa variance est la moyenne de X - ¯x¯. De façon plus précise, si X prend les valeurs (au point milieu) x1,x2, …,xk avec fréquence absolue n1,n2, …,nk et la population total est N, alors   
sigma2(x)=((x1 - ¯x¯)2n1+(x2 - ¯x¯)2n2+…+(xk - ¯x¯)2nk)/N Ex : p.3 NdeC

- On a des ‘’carrés’’ pour chaque terme pour qu’il soit positif et ainsi obtenir une mesure de la variance totale par rapport à ¯x¯.  
- Être sous ¯x¯ ajoute autant de variance qu’être en haut de ¯x¯ (si on est à la même distance)  
  
Définition 3 : Si X est une variable numérique sur un **échantillon** de taille N qu’on utilise pour approximer une population, et qu’on y a calculer une moyenne échantillonale ¯x¯, la variance échantillonale sera   
s2(x)=((x1 - ¯x¯)2n1+(x2 - ¯x¯)2n2+…+(xk - ¯x¯)2nk)/(N-1)  
- Le but de diviser par (N-1) est d’approximer plus fidèlement s2 sur la population  
- Sur la calculatrice s2: variance sur une population, sinon sigma2 : variance échantillonale

- Il est important de faire attention à la façon dont la question est posée pour s’assurer d’utiliser la bonne formule pour y répondre

19 septembre

-Par convention, on préfère « u » pour la moyenne d’une population et ¯x¯ pour la moyenne d’un échantillon, mais le calcul est le même.>  
-Par convention, on ne met pas d’unité à la variance.  
  
-L’écart type, notée sigma si population, s si échantillon, est la racine carrée de la variance  
-L’écart type essaie d’approximer la différence absolue entre la moyenne et une valeur typique.  
  
-Le coefficient de variation ou CV est défini comme l’écart type/moyenne x 100% est une mesure de à quelle point notre variable varie. Il aide à savoir à quel point nos valeurs sont-elles éparpillées.  
-CV n’a pas d’unité, c’est strictement une proportion  
-CV peut être plus grand que 100%  
-Dans ce cours, on considère une moyenne fiable, si le coefficient de variation est plus petit que 15%. Sinon on dit que la distribution est non-homogène. Cette limite peut varier d’un domaine à un autre.  
-On peut calculer le CV avec confiance si nos variables on les mêmes unités. Sinon c’est simplement une indication.

-Si x est une valeur que peut prendre une variable statistique de moyenne ¯x¯ et écart type sigma, la cote z de cette valeur est z=(x-moyenne)/sigma  
-La cote z n’a pas d’unité  
-La cote z d’une valeur essaie de quantifier à quel point une certaine valeur est aberrante dans la distribution.  
-Les valeurs habituelles sont entre -1 et 1 pour la cote z, il est par contre possible dans certaine situation exceptionnel que cela ne soit pas le cas.  
-Une valeur de cote z plus grande que un ou plus petite que -1 est dite anormale ou sont en évidence.  
-L’immense majorité des valeurs pour une distribution réaliste auront entre 3 et -3 de cote z.  
  
-Pour une distribution normale, (la plus fréquente) plus ou moins 68,26% des valeurs dans [-1,1] de cote z, plus ou moins 85,54% [-2,2] de cote z et plus ou moins 99,74 [-3,3] de cote z.

4 octobre 22

Si X est une variable aléatoire, la moyenne et l’écart type et ont une distribution normale. Alors x – la moyenne divisé par l’écart type est égal à Z. Donc la probabilité que X se trouve à l’intérieur de a;b, alors la probabilité que Z soit à l’intérieur sera de a-la moyenne divisé par l’écart type; b-la moyenne divisé par l’écart type.

Théorème : Si X est une variable statistique, de moyenne, de variance et que la moyenne d’un échantillon aléatoire de taille n, si n est plus grand que 30 alors la moyenne est N(moyenne;variance/par population)

La probabilité de 1 n’est pas égal que l’on est certain que cette chose se produise.

Théorème : Si A est qualitative, a une modalité de A, et qu’une proportion de p des individus ont la propriété A=a alors X = 1 si A = a ; 0 sinon. A la moyenne p, variance p(1-p).

1. P(1-P) = p-p2
2. P(1-P) est plus petit ou égal à ¼ de tout p
3. Étant donné que la variance est calculée directement avec p, il n’y a pas de distinction entre variance échantillonale et de population pour une proportion.

Remarque : Dans la réalité, on a les valeurs obtenues sur l’échantillon mais pas les valeurs réelles.

Les valeurs de l’échantillon sont une mesure, on va regarder leur marge d’erreur. À cause de l’élément arbitraire, on va fixer un niveau de confiance, plus ou moins à quel point on est confiant que la vraie valeur se trouve en dedans de notre marge d’erreur.

Remarque : En théorie, 1-a est la probabilité que, si quelqu’un refaisait la même expérience, il obtiendrait plus ou moins les mêmes résultats. En pratique, selon la discipline, il y a un niveau de confiance « standard » et l’intervalle de confiance représente une marge d’erreur similaire à celles des mesures physiques.

Formule : Échantillon taille plus grand ou égal à 30, moyenne, variance échantillonale, s2. Avec le niveau de confiance 1-a, intervalle de confiance voir formule notes première page Intervalles de Confiance.

Justification : On utilise les meilleurs approximations qu’on a : ¯x¯ , sigma2 = s2, alors ¯x¯ est à peu prêt N(¯x¯ , s2/n)

Sigma2 = s2/n

Sigma = Racine carré de s2/n

24 Octobre

- De façon plus général, si l’intervalle de confiance avec niveau de confiance 1-a est [a,b], la marge d’erreur ME = (b-a)/2  
- La marge d’erreur est l’écart maximale par rapport à la valeur centrale qui convient encore à notre résultat avec niveau de confiance 1-a.  
- La marge d’erreur résume en un seul chiffre la précision de notre étude.  
- Plus n est grand, plus la marge d’erreur est petite.  
  
- Si on a trouvé une proportion expérimentale ¯p¯, l’intervalle de confiance pour la proportion avec niveau de confiance 1-a est : voir notes de cours.  
  
  
- Si y=ax2+bx+c, le graphique sera : 1. Ouvert vers le haut si a>0 2. Ouvert vers le bas si a<0 3. Symétrique autour du max ou min  
- Si y=x-x2, coefficient de x2 = -1 ===== 1. Ouvert vers le bas 2. Possède un maximum  
- 0-02=1-12=0 parabole symétrique autour de x=1/2 donc ¯p¯-¯p¯2 <= (1/2)-(1/2)2 = ¼ pour tout ¯p¯  
  
- Attention : en pratique, on veut trouver le n le plus petit possible qui nous assure une marge d’erreur voulue.   
  
- Remarque : La marge d’erreur est inversement proportionnelle à la racine carrée de n donc pour couper la marge d’erreur de moitié, il faut quadrupler n. En pratique, il y a un point où cela ne vaut plus la peine. Pour un sondage d’opinion, n à peu près = 1000-2000 est généralement accepté.

31 Octobre 2022

Pour les fonctions qu’on a des chances de rencontrer dans ce cours : f(x) existe toujours sauf si 1. On aurait un zéro à un dénominateur 2. On aurait une racine paire d’un nombre négatif (racine carrée, racine quatrième, racine sixième, etc…) 3. Logarithme d’un nombre négatif ou nul  
  
But général :  
i) Régression : Avec deux variables liées, comment écrire l’une en fonction de l’autre aussi fidèlement que possible

Remarque : Prenez la droite x=2, ce sont tous les points pour lesquels x=2, soit une ligne verticale. Ceci n’est pourtant pas le graphique d’une fonction linéaire ni d’aucune fonction, car une fonction n’a pas deux points pour un même x.  
  
Remarque : En général, toute courbe dans le plan xy peut être le graphique d’une fonction si et seulement s’il y a au plus un y pour chaque x sur la courbe.

Remarque : Si vous voulez résoudre f(x) = ax+b, a et b inconnues, avec f(x1)=y, et f(x2) = y2. Soustraire marche toujours très bien puisque les b s’annulent.  
  
On peut trouver ‘’a’’ en une étape si on inscrit a=(y2-y1)/(x2-x1), puis résoudre pour ‘’b’’.

Remarque : Soit y=ax+b et y=Ax+B deux droites. Elles seront parallèles si et seulement si a=A et b n’est pas = B. Elles seront perpendiculaires si a=-1/A

7 Novembre 2022

-Si on a un exposant pair, il y aura toujours deux possibilités à sa racine. Si on veut la racine négative, on écrit -nracine de a (où n est pair).  
  
-Si n est un nombre naturel : (ab)n= ((ab)(ab)(ab)….) n fois. Cela pourrait aussi être écris tel : (a\*a\*a\*a…) \* (b\*b\*b\*b…) n fois.