Revision

Comment choisir un test statistique Questions a répondre:

- Combien des **échantillons (groups)** on a?
- Les échantillons sont:
  - **Dépendantes**/appariés?
    - (jumeaux/ données répétées/comparaison du partie gauche et droit d'un sujet/études appariées/deux tests diagnostiques ou méthodes de mesure qui observent les mêmes sujets)
  - **Indépendantes**?
- Quel est le **type des variables**?
- Combien des sujets?
  - Pour les **donness qualitatives**:
    - Tableau de contingence: **% des cellules theoriques < 5**?
- Quelle est la nature des donnes?
  - Pour les **données quantitatives**:
    - **Distribution normale?**
    - **Variances égales/ inégales**?

Precisions

- Ecart type = déviation standard
- Notations:
  - s – déviation standard,
  - S – déviation standard d' échantillonnage
    - $S = s * \sqrt{(n/(n-1))}$
- Echantillons:
  - Indépendants
  - Dépendants (appariées)
- Distribution gaussienne = normale

3

Les hypothèses statistiques

- La création des hypothèses:
- Certains **questions médicales** ont deux réponses opposées
  - on force beaucoup des questions dans ce format
- Les réponses correspond aux deux **modèles** possibles de la réalité
- Ce deux modèles sont nommées: **hypothèses**
  - L' hypothèse **nulle**: H0
    - **Il n'y a pas** un différence statistiquement significative entre 2/>=2 groups (ex. un traitement [ibuprofène vs. placebo] ou un facteur de risque [présent vs. absent]) en ce qui concerne la moyenne/ médiane/ variance/ fréquence … d'une caractéristique (ex résultat du traitement: la fréquence du guérison [oui vs. non] ou la moyenne de la température)
    - **Il n'y a pas** un relation/lien/association/dépendance/corrélation statistiquement significative entre 2 caractéristiques/variables: (ex. un Facteur de risque [présent vs. absent] – une maladie [oui vs. non] , ou un traitement [ibuprofène vs. placebo] – le résultat du traitement (ex: la fréquence du guérison [oui vs. non] ou la moyenne de la température))
  - L' hypothèse **alternative**: H1 (négation du H0)
    - **Il y a** un différence statistiquement significative entre 2/>=2 groups en ce qui concerne la moyenne/ médiane/ variance/ fréquence … d'une caractéristique
    - **Il y a** un relation/lien/association/dépendance/corrélation statistiquement significative entre 2 caractéristiques/variables

Les tests statistiques nous permet de faire la chois entre les deux possibilités (H1 / H0)

4

Les etapes d'un test statistique

- **Étape 1. Formuler** les **hypothèses statistiques**:
- **Étape 2.** Décider sur une **statistique approprie** du **test** (**paramètre du test**)
- **Étape 3.** Sélecter le **niveau de signification** - la valeur **alpha**. α = **0,05**
- **Étape 4.** Déterminer la ***valeur critique (v.c.)*** de la statistique du test
- on détermine - une région *critique* ou **région de *rejet*** *(RR)*
  - **Lois t, Z: RR = (-∞ , -v.c.] ∪ [v.c. , + ∞)**
  - **Lois F, Khi 2: RR = [v.c. , + ∞)**
- **Étape 5. Calculer la valeur de la statistique / paramètre du test**
- **Étape 6.** la **décision statistique** en **fonction de la région critique** :
  - Si **Z_0 est dans RR (région du rejet/critique) on rejet H_0 et on accepte H_1**
    - **Il y a une différence/ Il y a une relation –statistiquement significative**
  - Si **Z_0 est dans RnR (région de non rejet) on reste avec H_0 (on peut pas rejeter H_0)**
    - **On ne peut pas dire qu' il y a une différence/ il y a une relation statistiquement significative**
- **Étape 6'. La décision avec p-value**
  - **La décision avec p-value.**
  - Si **p-value < α (=0,05) on rejet H_0 => on accepte H_1**
    - **Il y a une différence/ Il y a une relation –statistiquement significative**
  - Si **p-value ≥ α (=0,05) => on reste avec H_0**
    - **On ne peut pas dire qu' Il y a une différence/ Il y a une relation –statistiquement significative**

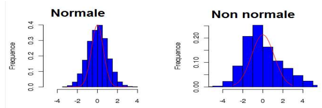
5

5

Vérification de la condition de normalité des données

Utilité:

- Importante pour appliquer des test paramétriques, avec condition de normalite:
  - Test Z pour les moyennes
  - Test t (Student)
  - Test ANOVA



Modalités de vérification (ici, <u>conditions de normalité</u>:

- des graphiques (les meilleures) modalités)
  - Histogramme (symétrique, comme un chapeau)
  - Boite a moustaches (symétrique autour de la médiane)
  - Le graphique des quantiles (voir diapositive suivant)
- des statistiques descriptives (pas très fiables)
  - Si la moyenne est ~= médiane
  - Si le coefficient de l'aplatissement ~=0 / apartient a [-1, 1] (kurtosis)
  - Si le coefficient de symétrie ~=0 / apartient a [-1, 1] (skewness)
- des tests de normalité: (ne sont pas recommandées)
  - Test de Kolmogorov-Smirnov (p < 0,05 – non normale, p > 0,05 ~normale)
  - Test de Shapiro-Wilk (p < 0,05 – non normale, p > 0,05 ~normale)

6

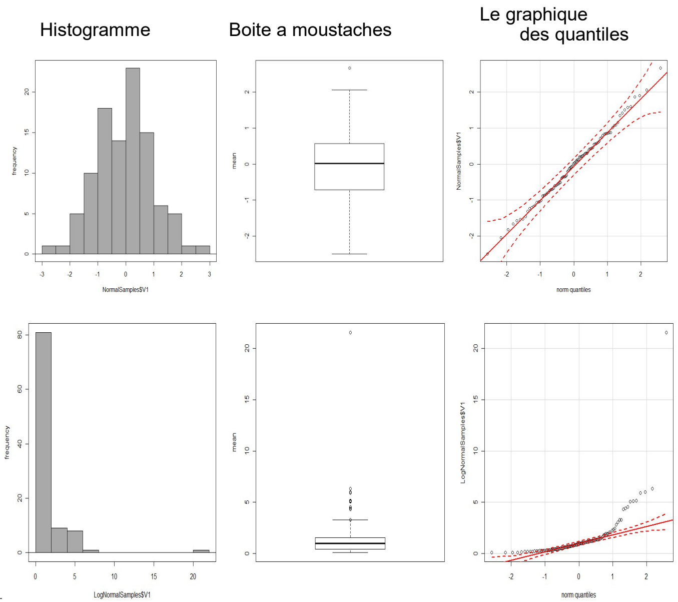
Comparaison des données normale/non normale distribuées

- **Normale**

moyenne ~= médiane
(=-0,03 =0,015)
c. asymétrie = 0,11
apartient a [-1, 1], ~= 0
c. aplatissement = -0,09
apartient a [-1, 1] ~= 0
Shapiro-Wilk test
p= 0,99 > 0,05

- **Non normale**

moyenne <> médiane
(=1.57 =0,98)
c. asymétrie = 5,59
> 1, <> 0
c. aplatissement = 40,63
> 1, <> 0
Shapiro-Wilk test
p=~0 < 0,05



La normalité des données

- Test de normalité des données :
  - Test de Kolmogorov-Smirnov
  - Si < 50 sujets le test Shapiro-Wilk
- H0 = aucune différence statistiquement significative entre la distribution observée et la distribution normale
- H1 = aucune différence statistiquement significative entre la distribution observée et la distribution normale
- p <0,05 on rejeté l'hypothèse nulle, les données ne sont pas normalement distribués
- p >0,05 on ne rejeté pas l'hypothèse nulle, on n'a pas des motifs pour considérer les données anormales – on peut considérer les données normale distribuées (dans certaines conditions)

Récapitulatif des tests utilises

Tests pour variables quantitatives - comparer la moyenne des deux échantillons

Type variable	Nb sujets	Nature des données	Statistique comparée	Test utilise	Paramètre du test	Région du rejet – test bidirection
Deux échantillons indépendants Pour l'examen écrit je ne demande pas les choses en gris						
Quantitative	n_1, n_2 ≥ 30	Normale distribuées, Variances dans la population connues inégales	Différence des moyennes	Test Z pour la différence entre les moyennes	$Z = \frac{m_1 - m_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	(-∞, -v.c.] U [v.c., +∞)
	n_1, n_2 ≥ 30	Normale distribuées, Variances dans la population connues égales	Différence des moyennes	Test Z pour la différence entre les moyennes	$Z = \frac{m_1 - m_2}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}}$	(-∞, -v.c.] U [v.c., +∞)
	n_1, n_2 ≥ ou < 30	Normale distribuées, Variances dans la population inconnues inégales	Différence des moyennes	Test t (Student) n_1 + n_2 - 2 d.d.l.	$t = \frac{m_1 - m_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$	(-∞, -v.c.] U [v.c., +∞)
	n_1, n_2 ≥ ou < 30	Normale distribuées, Variances dans la population inconnues égales	Différence des moyennes	Test t (Student) n_1 + n_2 - 2 d.d.l.	$t = \frac{m_1 - m_2}{\sqrt{\frac{S^2}{n_1} + \frac{S^2}{n_2}}}$ $S^2 = \frac{(n_1 - 1) \times S_1^2 + (n_2 - 1) \times S_2^2}{n_1 + n_2 - 2}$	(-∞, -v.c.] U [v.c., +∞)
Deux échantillons dépendants (appariées)						
Quantitative	$n_1 = n_2$ ≥ ou < 30	Normale distribuées,	Moyenne des différences	Test t (Student) n - 1 d.d.l.	$t = \frac{\bar{d}}{\frac{S}{\sqrt{n}}}$	(-∞, -v.c.] U [v.c., +∞)

(ou n, n_1, n_2 - nombre des sujets; m, m_1, m_2 – moyennes; s, s_1, s_2 – déviations standard descriptive de l'échantillon; S, S_1, S_2 – déviation standard d'échantillonage; $S = \sqrt{\frac{n}{n-1}}s$; $s = \sqrt{\frac{n-1}{n}}S$; σ, σ_1, σ_2 - déviation standard dans la population; pour α=0,05, Z_α=1,96; d.d.l. – dégrées de liberté; v.c. – valeur critique)

Récapitulatif des tests utilises

Tests statistiques pour **comparer deux échantillons, variables quantitatives non normale distribuées**

Type variable	Nature des données	Statistique comparée	Test utilise	Paramètre du test	Région du rejet
Tests statistiques pour comparer deux échantillons indépendants					
Quantitative ou qualitative ordinale	Données non normale distribuées	~médiane (si les distributions sont similaires) - la distribution des rangs /moyenne des rangs	Test Mann Whitney U Etapes: Tableaux avec **toutes les valeurs** des deux échantillons **Ordonnées** en ascendant **Numéroter** du plus petit a le plus grand **Donner un rang**. (Valeurs identiques, le rang = moyenne de la numérotation) **Calculez la somme des rangs** pour group A et B	$obsSR_a$ somme rangs group a, $obsSR_b$ somme rangs group b, $U_b = n_a \times n_b + n_b \times \frac{(n_b+1)}{2} - obsSRb$ $U_a = n_a \times n_b + n_a \times \frac{(n_a+1)}{2} - obsSR_a$	Min (Ua , Ub) <= v.c.
Tests statistiques pour deux échantillons <u>dépendants/ apariees</u>					
Quantitative ou qualitative ordinale	Données non normale distribuées	~médiane (si les distributions sont similaires) - la distribution des rangs	Test Wilcoxon pour échantillons appariées Etapes: Calculer la **différence** des paires **Ignorer les zéros** **Ignorer les signes** **Numéroter** du plus petit a le plus grand **Donner le rang** 1, 2, 3… (Valeurs identiques, le rang = moyenne de la numérotation) **Calculez la somme des rangs positifs** (W+), et puis **négatifs** (W-).	W_- somme rangs négatifs W_+ somme rangs positifs	Min (W_- , W_+) <= v.c.

v.c. – valeur critique. na – nombre sujets dans group a, nb - nombre sujets dans group b. Min – le minimum entre les deux valeurs.

Récapitulatif des tests utilises

Type variable	Nb sujets	Nature des données	Statistique comparée	Test utilise	Paramètre du test	Région du rejet
Tests statistiques pour deux échantillons indépendants, variables qualitatives dichotomiques						
Qualitative dichotomique	$n_1p_1>10$, $n_2p_2>10$, $n_1(1-p_1)>10$, $n_2(1-p_2)>10$		fréquences	Test Z * Equivalent au test Chi carrée ci dessous (Chi carrée est recommandé):	$Z=\frac{p_1-p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1}-\frac{1}{n_2}\right)}}$ $p=\frac{p_1n_1+p_2n_2}{n_1+n_2}$	$\left(-\infty,-Z_{\frac{\alpha}{2}}\right]\cup\left[Z_{\frac{\alpha}{2}},+\infty\right)$
	<20% cellules du tableau théorétique sont <5		fréquences	Test Chi carrée * v=(L-1)*(C-1) = 1 d.d.l.	$\chi^2=\sum_{i=1}^{L*C}\frac{\left(f_i^o-f_i^t\right)^2}{f_i^t}$	$\left[\chi^2_{v\alpha},+\infty\right)$
	>20% cellules du tableau théorétique sont <5		fréquences	Test exact Fisher	- (test non paramétrique)	-
Tests statistiques pour deux échantillons dépendants , variables qualitatives dichotomiques						
Qualitative dichotomique	b+c>25		fréquences	Test Chi carrée v=1 d.d.l.	$\chi^2_{1ddl}=\frac{\left(\lvert b-c\rvert-0{,}5\right)^2}{b+c}$	$\left[\chi^2_{v\alpha},+\infty\right)$

* On préfère pour ce cours (et l' examen) le test Chi carrée pour comparer les fréquences, au lieu du test Z pour comparer les fréquences

p_1, p_2 – fréquences; n_1, n_2 – nombre des sujets; L et C – nombres des lignes et des colonnes dans le tableau de contingence, f^o – fréquence observée, f^t – fréquence théorétique; d.d.l. – dégrées de liberté;

Pour l'examen écrit je ne demande pas les choses en gris

Récapitulatif des tests utilises

Type variable	Nb sujets	Nature des données	Statistique comparée	Test utilise	Paramètre du test	Région du rejet
Tests statistiques pour plus des deux échantillons (groups) indépendants						
Qualitative	<20% cellules du tableau théorétique sont <5		Fréquence	Test Chi carrée v=(L-1)*(C-1) d.d.l.	$\chi^2=\sum_{i=1}^{L*C}\frac{\left(f_i^o-f_i^t\right)^2}{f_i^t}$	$\left[\chi^2_{v\alpha},+\infty\right)$
	>20% cellules du tableau théorétique sont <5		fréquence	Test exact Fisher	-	-
Quantitative		Normale distribuées, Variance des échantillons égalés	moyenne	Test ANOVA v1 = nb. groups – 1 v2 = nb. obs. – nb groups	$F=\frac{MCG}{MCE}$	$\left[F_{v_1,v_2,\alpha},+\infty\right)$
		Normale distribuées, Variance des échantillons inégalés	moyenne	ANOVA de Welch ou Brown Forsyth		
		Non normale distribuées	médiane	test Kruskal Wallis	-	-

L et C – nombres des lignes et des colonnes dans le tableau de contingence, f^o – fréquence observée, f^t – fréquence théorétique; d.d.l. – dégrées de liberté;

MCEG = [n1 * (m1-mt)^2 + n2 * (m2-mt)^2 + n3 * (m3-mt)^2 + …] / (p-1)
MCDG = [(n1-1) * DS1^2 + (n2-1) * DS2^2 + (n3-1) * DS3^2 + …] / (n-p)
n=nombre total d'observations, n1,n2, n3 – le nombre d'observations par group, p – nombre des groups
mt – la moyenne des toutes les observations, m1, m2, m3… les moyennes par groups, DS1, DS2, DS3, … déviations standard d' échantillonnage des groups

Récapitulatif des tests utilises

Tests statistiques pour comparer les variances entre deux échantillons						
Type variable	Nb sujets	Nature des données	Statistique comparée	Test utilise	Paramètre du test	Région du rejet
Quantitative	Données normale distribuées		variances	Test F, $v_1 = n_1$ d.d.l. $v_2 = n_2$ d.d.l.	$F = \begin{cases} \frac{S_2^2}{S_1^2}, pour \cdot S_2^2 \rangle S_1^2 \\ \frac{S_1^2}{S_2^2}, pour \cdot S_1^2 \rangle S_2^2 \end{cases}$	$\left[F_{v_1, v_2, \alpha}, +\infty\right)$
Tests statistiques pour comparer les variances entre > deux échantillons						
Quantitative			variances	Test Bartlett ou Test Levene		

(ou n_1, n_2 - nombre des sujets; m_1, m_2 – moyennes; s_1, s_2 – déviations standard descriptive de l'échantillon; S_1, S_2 – déviation standard d'échantillonage; $S = \sqrt{\frac{n}{n-1}}s$; $s = \sqrt{\frac{n-1}{n}}S$; d.d.l. – dégrées de liberté)

Equivalences entre tests paramétriques et non paramétriques

Données	Nombre échantillons	Tests paramétriques	Tests non parametriques
qualitatives		Chi deux	exact Fisher
Quantitatives (ou qualitatives ordinales)	2 indépendants	Student (t) pour échantillons indépendants	Mann Whitney U (Wilcoxon somme des rangs Mann Whitney Wilcoxon)
	2 appariées (dépendants)	Student (t) pour échantillons appariées	Wilcoxon rangs signées (Wilcoxon pour échantillons appariées)
	> 2 indépendants	ANOVA (pour variances égales) ou ANOVA de Welch ou Brown Forsyth (pour variances inégales)	Kruskal Wallis

Pour données **normale** distribuées

Pour données **non normale** distribuées

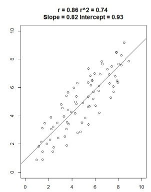
Intervalles de confiance

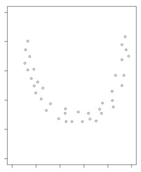
Type variable	Nombre echantillons	Estimateur ponctuel	Conditions application	Formule
qualitative	une	fréquence: f	grands échantillons: nf ≥ 10 et nq ≥ 10	$\left(f - Z_{\frac{1-\alpha}{2}} ES; f + Z_{\frac{1-\alpha}{2}} ES\right)$ $ES = \sqrt{\frac{f(1-f)}{n}}$
	une	fréquence: f	petits échantillons: nf < 10 ou nq <10	on ne va pas calculée
	deux	différence entre les fréquences: $f_1 - f_2$	grands échantillons: $f_1n_1 \geq 10$, $(1-f_1)n_1 \geq 10$, $f_2n_2 \geq 10$, $(1-f_2)n_2 \geq 10$	$ES = \sqrt{\frac{f_1 \times (1-f_1)}{n_1} + \frac{f_2 \times (1-f_2)}{n_2}}$ $\left((f_1 - f_2) - Z_{\frac{1-\alpha}{2}} ES; (f_1 - f_2) + Z_{\frac{1-\alpha}{2}} ES\right)$
	deux	différence entre les fréquences:	petits échantillons: np < 10 ou nq <10	on va pas calculée
quantitative	un	moyenne: m	grands échantillons: n ≥ 30, σ - connue	$\left[m - Z_{\frac{1-\alpha}{2}} ES; m + Z_{\frac{1-\alpha}{2}} ES\right]$ $ES = \frac{\sigma}{\sqrt{n}}$
	un	moyenne: m	grands échantillons: n ≥ 30, σ – non connue	$\left[m - t_{n-1,\frac{1-\alpha}{2}} ES; m + t_{n-1,\frac{1-\alpha}{2}} ES\right]$ $ES = \frac{S}{\sqrt{n}}$
	un	moyenne: m	petits échantillons: n < 30, σ – non connue	$\left[m - t_{n-1,\frac{1-\alpha}{2}} ES; m + t_{n-1,\frac{1-\alpha}{2}} ES\right]$ $ES = \frac{S}{\sqrt{n}}$
	deux	différence entre les moyennes: $m_1 - m_2$	variances égales	$ES = \sqrt{\frac{(n_1-1) \times s_1^2 + (n_2-1) \times s_2^2}{n_1 + n_2 - 2}}$ $\left((m_1 - m_2) - t_{n_1+n_2-2,\frac{1-\alpha}{2}} ES; (m_1 - m_2) + t_{n_1+n_2-2,\frac{1-\alpha}{2}} ES\right)$

(ou n, n1, n2 - nombre des sujets; f, f1, f2 - fréquence observée; q = 1 – f; m, m1, m2 – moyennes; s, s1, s2 – déviations standard descriptive de l'échantillon; S – déviation standard d'échantillonage; $S = \sqrt{\frac{n}{n\_1}} s$; $s = \sqrt{\frac{n\_1}{n}} S$; σ - déviation standard dans la population; ES=erreur standard; pour α=0,05, Zα=1,96)

Evaluation graphique : diagramme de dispersion (nuage des points/scatter)
Evaluer la linearite, et l'importance de la correlation:

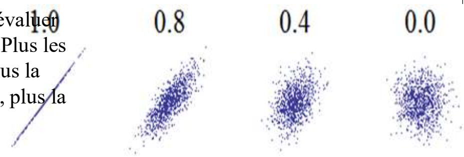
Si les points semble suggérer un droite – la relation est peut être linéaire





Si les points semble suggérer des tendances qui ne sont pas linéaires, la relation est peut être non linéaire (exponentielle, quadratique)

Si la relation est plus probable linéaire, on peut évaluer d'une manière subjective la corrélation linéaire. Plus les points se rapprochent a un droite de tendance, plus la corrélation est forte, plus les points sont distants, plus la corrélation est faible

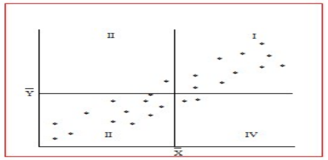


Evaluation graphique : diagramme de dispersion (nuage des points/scatter)

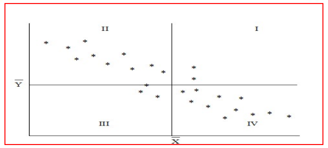
But : évaluation visuelle de la relation entre deux variables quantitatifs

L'utilisation des cadrans pour identifier **la tendance/sens/direction** directe/inversement proportionnelle :

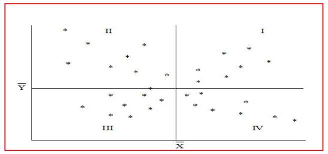
i) presque toutes les points sont dans les cadrans I et III ⇒ tendance croissante/ pente ascendante/ pente positive/ lien (direct) proportionnel



ii) presque toutes les points sont dans les cadrans II et IV ⇒ tendance décroissante / pente descendante/ pente négative/ lien inversement proportionnelle



iii) les points sont distribues uniformément dans tous les cadrans ⇒aucune tendance



Corrélations

Type des variables	Nature des données	Coefficient de correlation	Formule du coefficient
quantitative	normale distribuées	Pearson (r)	$COV(X,Y)=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})$ $r=\frac{COV(X,Y)}{S_X \cdot S_Y}$
quantitative	non normale distribuées	Spearman (ρ - rho)	$\rho=r_s=1-\frac{6\sum_{i=1}^{n}d_i^2}{n(n^2-1)}, ou\ d_i=R_{x_i}-R_{y_i}$
qualitative ordinales	-	Spearman (ρ - rho)	$\rho=r_s=1-\frac{6\sum_{i=1}^{n}d_i^2}{n(n^2-1)}, ou\ d_i=R_{x_i}-R_{y_i}$

Xi , Yi – sont les valeurs des deux séries des données. $\bar{X}$ et $\bar{Y}$ sont les moyennes des deux séries. Rxi et Ryi sont les rangs des valeurs Xi et Yi après leurs rangement dans ordre croissante. n – nombre des observations. Sx et Sy sont les déviations standard d'échantillonnage. COV(X,Y) – la covariance

Covariance COV(X,Y) : **Corrélation linéaire Pearson - interprétations**

- >0 tendance croissante/ pente ascendante/ lien direct proportionnel/ covariance positive
- <0 tendance décroissante/ pente descendante/ lien inversement proportionnel/ covariance négative
- $\cong 0 \Rightarrow$ aucune tendance

r (coefficient de corrélation Pearson):

montre la direction et l' intensité de la corrélation;

Interprétation du <u>direction/sens/tendance</u>:

- >0 tendance croissante/ pente ascendante/ lien direct proportionnel/ corrélation positive
- <0 tendance décroissante/ pente descendante/ lien inversement proportionnel/ corrélation négative
- $\cong 0 \Rightarrow$ aucune tendance

- plus **r** ou COV(X,Y) est grand (en valeur absolue) plus la relation est forte
- plus **r** ou COV(X,Y) est proche du 0, plus la relation est faible

19

Le coefficient Pearson - interprétation

Interprétation de l'intensité/force/degré/importance de la corrélation linéaire avec les règles empiriques de Colton [Colton T. Statistics in Medicine. Little Brown and Company, New York, NY 1974] **(on préfère le mot corrélation ici,** même si association/lien/relation peut être utilise**)**

(-0.25 et 0,25)

=> une relation **négligeable** ou **aucune** corrélation linéaire entre les variables

[0.25 et 0.50) ou [-0.25 et -0.50)

=> un degré de corrélation **faible/acceptable**

[0.50 et 0.75) ou [-0.50 et -0.75)

=> un degré de corrélation **modérée à bonne**

[0.75 et 1] ou [-0.75 et -1]

=> une **très bonne à excellente** corrélation

Il y a autre divisions possibles aussi.

Ces règles doit être utilisée avec soins. Elle sont pour donner une idée, mais pour chaque problème, l'intensité de la relation est relative au domaine. Pour certain situations les valeurs en dessous de 0,8 peut être faibles.

20

Régression linéaire simple

- **Interprétation**
- La droite de régression Y(X): $Y = b_0 + b_1 X$

b0 = est l'ordonnée à l'origine – la valeur du Y quand X est égal a 0 (d'habitude cette information n'est pas utile pour les médecins, elle présente une situation qui en réalité est impossible)

b1 = la pente de la droite de régression.

Interprétation de b1 - du coefficient de la variable X

chaque unité de mesure de la variable indépendante - X en plus augmente en moyenne la variable dépendante – Y avec la valeur du coefficient de la variable indépendante X – b1

21

21

Régression linéaire multiple - Quantification de l'importance de la relation pour plusieurs variables

- L' équation du **régression linéaire multiple**:
- Variable dépendante= coefficient_1 * variable_1 + coefficient_2 * variable_2 + … + coefficient_n * variable_n + coefficient_0
- Ex: triglycérides (mg/dL) = 23,10 * obésité (oui/non) + 1,14 * cholestérol (mg/dL)
- **L' interprétation du coefficient ajusté (adjusted – en anglais)** pour des variables **Qualitatives dichotomiques** (ex. obésité):
  - l'augmentation de la variable dépendante – les triglycérides en moyenne (ici il est de 23,1 mg/dL) pour ceux qui ont le facteur présent (être obese - la variable indépendante) comparée a ceux qui n'ont pas le facteur (ne sont pas obeses), si on tiennent les autres variables constantes / si on ajuste les autres variables/ si on contrôle les autres variables) (ici – le cholestérol)
  - ceux qui ont le facteur présent (être obese - la variable indépendante) ont la variable dépendante – les triglycérides en moyenne plus grand avec 23,1 mg/dL comparée a ceux qui n'ont pas le facteur (ne sont pas obeses), si on tiennent les autres variables constantes / si on ajuste les autres variables/ si on contrôle les autres variables) (ici – le cholestérol)

22

22

Variables aléatoires discrètes
Esperance, variance et écart type

- **Esperance mathématique** ou la moyenne théorique d'une v.a. X

$$E(X)=\sum_{i=1}^{n} x_i \Pr(X=x_i) \sim \boxed{\bar{x}=\frac{1}{n}\sum_{i=1}^{n} x_i}$$

Exemple: avec le traitement

- X: 0 1 2 3 4

Pr 0,008 0,076 0,256 0,411 0,24

E(X) = 0 x 0,008 + 1 x 0,076+2 x 0,265 + 3 x 0,411 + 4 x 0,24 = 1,31

- **Variance** de X:

$$V(X)=\sum_{i=1}^{n}\left[x_i - E(X)\right]^2 * \Pr(x_i) \sim \boxed{\sigma^2=\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}}$$

- **Ecart-type** (deviation standard) de X

23

$$\sigma(X)=\sqrt{V(X)}=V(X)^{1/2}$$

Tableau récapitulatif

Evénements	*Définitions*	*Notations*	*Calcul des probabilités*
Evénement contraire d'un événement A	l'événement constitué par tous les événements élémentaires qui ne sont pas dans A.	$\bar{A}$	Propriété : $\Pr(\bar{A}) = 1 - \Pr(A)$
Evénement "A et B" ou **intersection de A et B**	l'événement "A et B" est constitué par tous les événements élémentaires se trouvant à la fois dans A et dans B.	$A \cap B$	
Evénement "A ou B" ou **réunion de A et B**	l'événement "A ou B" est constitué par tous les événements élémentaires se trouvant dans l'un au moins des événements A ou B	$A \cup B$	Propriété : $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$
Evénements incompatibles	ils ne peuvent pas être réalisés simultanément.	$A \cap B = \varnothing$	$\Pr(A \cap B) = 0$
Evénements indépendants	Deux événements sont indépendants si la survenance d'un événement n'affecte pas l'apparition d'un deuxième événement.		$\Pr(A \cap B) = \Pr(A) \times \Pr(B)$

Probabilité conditionnelle

- A et B évènements, $P(A) \neq 0$.

❑ Notation: $Pr(B \mid A)$ **:** probabilité conditionnelle de B, sachant que l'événement A est réalisé

❑ **Formule de calcul: $Pr(B \mid A) = Pr(A \cap B)/Pr(A)$**

Ex.

1) la probabilité d'avoir un cancer colorectal sachant que le test Hémocult est positif, est une probabilité conditionnelle:

Pr(cancer colorectal | test Hémocult Positif)

2) la probabilité d'avoir un cancer oral sachant que le test de bleuissement avec toluidine est positif

Pr(cancer oral | bleuissement avec toluidine Positif)

3) la probabilité d'un homme d'avoir

PAS > 140 mmHg: Pr(PAS > 140 | Hommme)

Rajmohan M, Rao UK, Joshua E, Rajasekaran ST, Kannan R. Assessment of oral mucosa in normal, precancer and cancer using chemiluminescent illumination, toluidine blue supravital staining and oral exfoliative cytology. J Oral Maxillofac Pathol. 2012 Sep;16(3):325-9. http://dx.doi.org/10.4103/0973-029X.102476

25

Probabilité conditionnelle et l'indépendance

❑ A , B évènements **indépendantes**:

$$Pr(B \mid A) = Pr(B) = Pr(B \mid \text{non } A)$$

❑ A et B évènements **dépendantes**:

$$Pr(B \mid A) \neq Pr(B) \neq Pr(B \mid \text{non } A)$$

$$Pr(A \cap B) \neq Pr(A) \times Pr(B)$$

26

Probabilité conditionnelle - applications

Ex.: enquête des possibles facteurs de risque du cancer de poumon

Dans un échantillon de 2000 sujets, on a 1000 fumeurs parmi lesquelles 130 sujets souffrent de cancer du poumon et 1000 non fumeurs parmi lesquelles 10 sujets souffrent de cancer du poumon

Le risque relatif d'avoir le cancer de poumon = ?

Solution: *on considère les événements*

A= {sujet fumeur} et B= {sujet atteints de cancer du poumon}

$\Pr(B \mid A) = 130/1000 = 0{,}130$ $\qquad \Pr(B \mid \overline{A}) = 10/1000 = 0{,}010$

$$RR = \frac{\Pr(B \mid A)}{\Pr(B \mid \overline{A})} = \frac{0{,}130}{0{,}010} = 13$$

=>un sujet qui est fumeur a un risque d'avoir le cancer du poumon de 13 fois plus grand qu'un sujet qui n'est pas fumeur

Probabilité conditionnelle - applications

	M+	M-	Total
T+	a	b	a+b
T-	c	d	c+d
Total	a+c	b+d	n

Se = $\Pr(T^+ \mid M^+) = a / (a + c)$

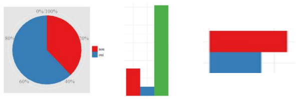
Sp = $\Pr(T^- \mid M^-) = d / (b + d)$

VPP = $\Pr(M^+ \mid T^+) = a / (a + b)$

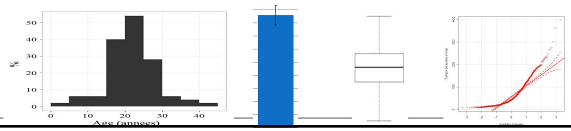
VPN = $\Pr(M^- \mid T^-) = d / (c + d)$

Le chois du type du graphique en fonction des types des variables et but

- Pour faire la **choix**, comptez **combien des variables** sont et quel **est le type.**
- **Description d'une seule variable**
  - *Qualitative*
    - camembert (sectoriel – **Pie**)
    - **Column** (si les noms des catégories ne sont pas très longues)
    - **Bar** (si les noms des catégories sont très longues)

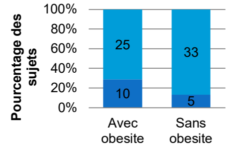


  - *Quantitative*
    - **Histograme**, graphique des moyennes, box and wiskers (boite a moustaches), graphique des quantiles

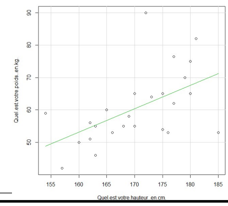


Le chois du type du graphique en fonction des types des variables et but

- **La relation entre deux variables**
  - *Qualitative*
    - **Column** (Clustered Column/ Stacked Column/ 100% Stacked column), ou **Bar** (Clustered Bar / Stacked Bar / 100% Stacked Bar)

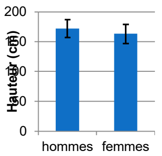


  - *Quantitative*
    - **Scatter** (nouage des points)

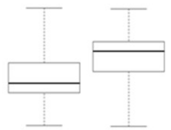


Le chois du type du graphique en fonction des types des variables et but

- **La relation entre deux variables**
  - ***Une variable quantitative en fonction d'une variable qualitative***
    - **Si les donnees sont normale distribuees**
      - Graphique des moyennes (avec deviation standard)
      - Graphique Colonnes avec barre d' erreur



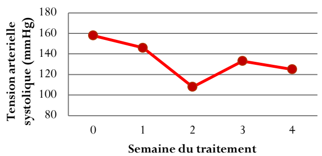
    - **Si les donnees sont normale distribuees**
      - Graphique box plot ou wiskers/boite a moustaches (boite a moustaches)



31

Le chois du type du graphique en fonction des types des variables et but

- **L'évolution dans le temps d'une variable qualitative ou quantitative**
  - **Line** (Ligne)



- **La relation entre trois variables quantitatives**
  - **Bubble** (nouage des sphères)
  - Nouage des points tridimensionnel
- **Une variable qualitative en fonction des intervalles d'une variable quantitative**
  - **Area** (Surface)

32

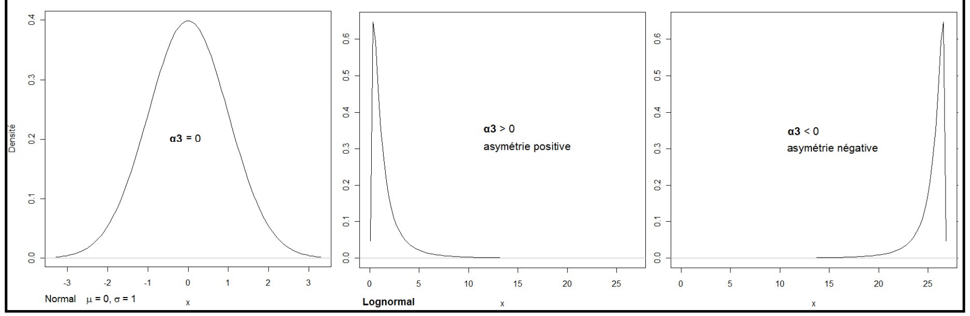
Mesures de symétrie: (skewness)

Coefficient d'asymétrie (α3):

degré d'asymétrie d'une distribution
la direction de cette asymétrie (positive ou négative);

α3 ≈ 0 => une distribution symétrique.
α3 > 0 => distribution est plus allongée vers la droite – asymétrie positive
α3 < 0 => distribution est plus allongée vers la gauche – asymétrie négative

α3 (-0,5 – 0,5) approximative symétrique
α3 (-1 – -0,5) ou (0,5 – 1) modérée asymétrique
α3 < -1 ou >1 asymétrie importante



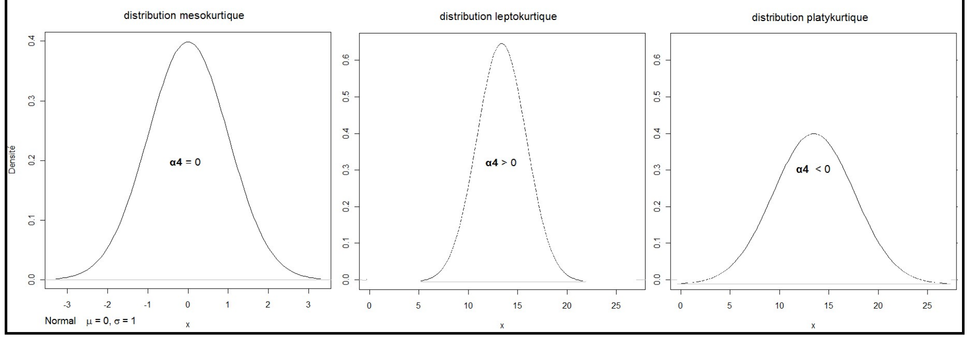
33

Le coefficient d'aplatissement (Kurtosis)

Le coefficient d'aplatissement (α4):

l'angle de la courbe du milieu d'une distribution
par rapport a une distribution normale (gaussienne)

- **α4** ≈ 0 => l'angle normal => distribution mesokurtique
- **α4** > 0 => l'angle aigu => distribution leptokurtique - centre élevée
- **α4** < 0 => la pente aplaté => distribution platykurtique – centre plus bas



34

Statistique descriptive

Mesures de tendance centrale:	Mesures de dispersion:
✓ **Moyenne** ✓ **Médiane** ✓ Mode	✓ Amplitude (entendue) ✓ **Intervalle interquartile** ✓moyenne des écarts de la moyenne ✓moyenne des écarts de la médiane ✓ Variance ✓ **Déviation standard** (**écart-type**) ✓ Coefficient de variation ✓ Erreur standard
Mesures de symétrie/aplatissement: ✓ Coefficient **d'asymétrie** (skewness) ✓ Coefficient **d'aplatissement** (Kurtosis)	**Mesures de position:** ✓ **Quartiles** ✓ Déciles ✓ Percentiles

35

Bon success!!!

36

36