Covariance & Correlation

Stats with R - TD 2

Jeanne Charoy - Automne 2021

Rappel: Les bases sur R

Trouver des infos sur les fonctions:

- help([nom de la fonction]): help(mean), help(c), help(plot)....
- ?[nom de la fonction]: ?mean, ?c, ?barplot….
- Google: “how to calculate mean in R”, “how to make a bar plot in R”.....

Debugging:

- Toujours commencer par checker la syntaxe, c’est souvent ca le probleme
- Copy-paste le message d’erreur dans google
- Utilise des print (a un niveau plus avance)

Ce qu’on a vu la derniere fois

- Comment creer un projet + script dans R
- Comment telecharger des packages
- Fonctions de bases: c(), mean(), median(), sd(), …
- Assigner des valeurs a des objets
- Lire un fichier csv et le télécharger dans l'environnement R
- Manipuler un dataframe:
  - Creer nouvelle colonne
  - Selectionner un subset de la dataframe

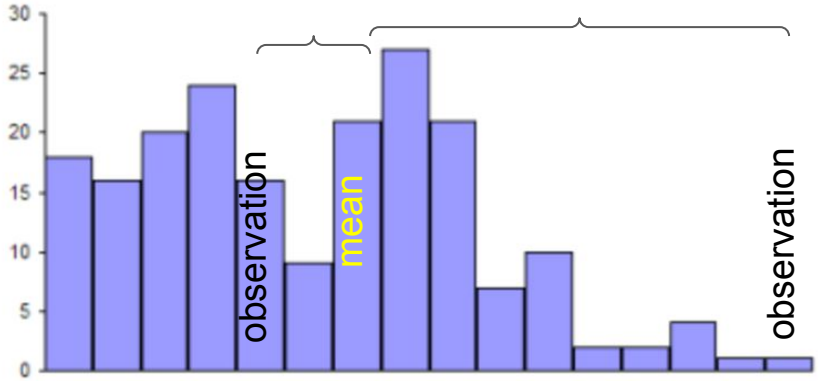
Aujourd’hui

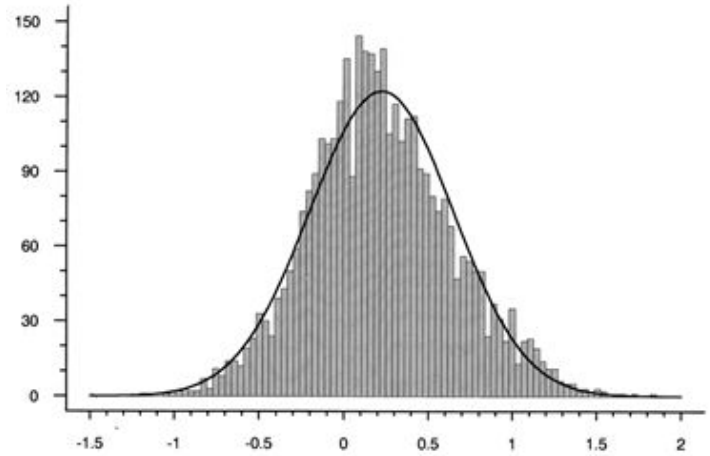
- Statistiques univaries (i.e., une seule variable dont on observe la dispersion)
- Standardization
- Statistiques bivaries (i.e., deux variables qui bougent ensemble)
- Covariance
- Correlation

Statistiques univaries

Deux concepts important:

- La tendance centrale (e.g., mean, median, mode): le point centrale d'une distribution univarie
- La dispersion: l'organisation des observations autour de la tendance centrale





Dispersion

Mesurer la tendance centrale est plutot facile (mean, etc…), mais comment representer la dispersion des observations avec un seul index numerique?

Ca depend en partie du point centrale que l’on choisit.

En general on utilise la moyenne arithmetique

Dispersion

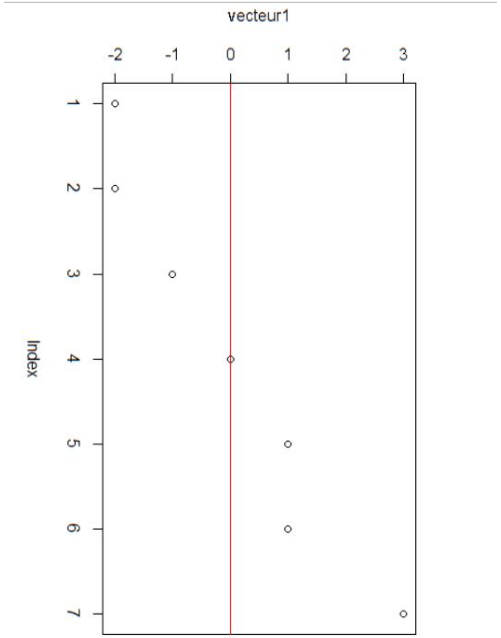
Imaginons sept observations: vecteur1 < - c(-2, -2, -1, 0, 1, 1, 3)

Quelle est la moyenne?

Dessinons un scatterplot pour visualiser les donnees:

plot(vecteur1)

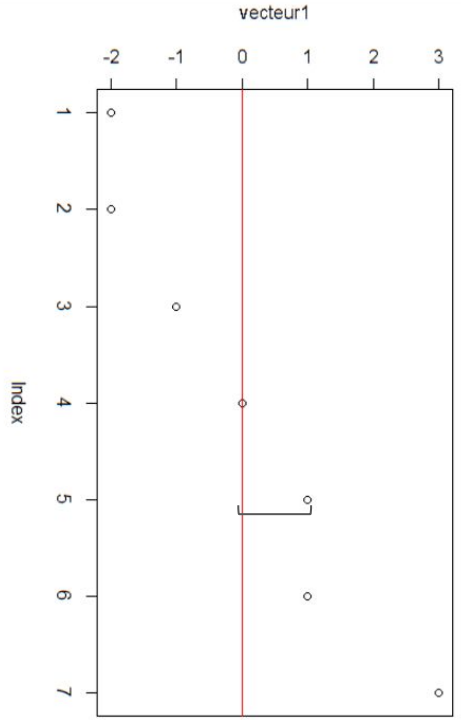
abline(h = mean(vecteur1), col=”red”)



Dispersion

On peut calculer **l'ecart** de chaque observation par rapport a la moyenne.

Valeur	Ecart a la moyenne
-2	-2
-2	-2
-1	-1
0	0
1	1
1	1
3	3



Dispersion

Comment caractériser la dispersion de ces valeurs? C'est tentant de faire la moyenne des ecarts: $\Sigma X_{dev} / N$

Dans R: ecarts <- c(-2, -2, -1, 0, 1, 1, 3)

mean(ecarts) = ?

Dispersion

Comment caractériser la dispersion de ces valeurs? C'est tentant de faire la moyenne des ecarts: $\Sigma X_{dev} / N$

Dans R: ecarts <- c(-2, -2, -1, 0, 1, 1, 3)

mean(ecarts) = ?

Par nature, la moyenne des ecarts type est toujours egale a zero. Il faut donc un moyen de calculer la moyenne des ecarts sans que les valeurs + et - ne s'annulent:

- Valeur absolue: mais mathematiquement ca pose des problemes plus tard
- Quoi d'autre?

Variance

Dans R: Comment creer un nouveau vecteur “ecart2” qui contient le carree des ecart?

mean(ecart2) = ?

Ceci nous donne la **variance** = la moyenne du carree des ecarts a la moyenne

$$s^2 = \frac{\sum (X - \overline{X})^2}{N-1} = \frac{\sum (X - \overline{X})(X - \overline{X})}{N-1}$$

Ecart type

L’ecart type est la racine carree de la variance:

ecartType <- sqrt(ecart2)

Pourquoi ne pas utiliser la variance?

- On a eleve les ecarts a la puissance 2
- La vairance est donc mesuree en unite au carree et pas dans l’unite de depart
- L’ecart type permet d’exprimer la dispersion des observations dans l’unite des valeurs de depart (e.g., cm, kg, seconds…)

Ecart type + variable centree reduite

On peut utiliser l'ecart type pour creer des variable centree reduite, aussi appele z scores (et stqndardized scores en anglais)

Les z scores sont les valeurs de depart, simplement transformees de maniere a ce que la moyenne soit 0 et l'ecart type 1.

Les z score expriment la distance entre une observation et la moyenne en term du nombre d'ecart type.

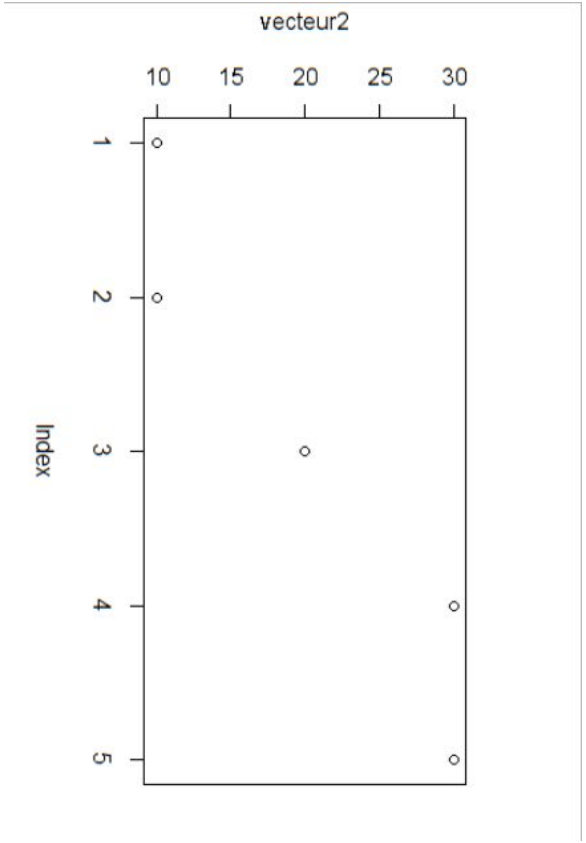
$$z = \frac{X - \overline{X}}{SD}$$

Z scores

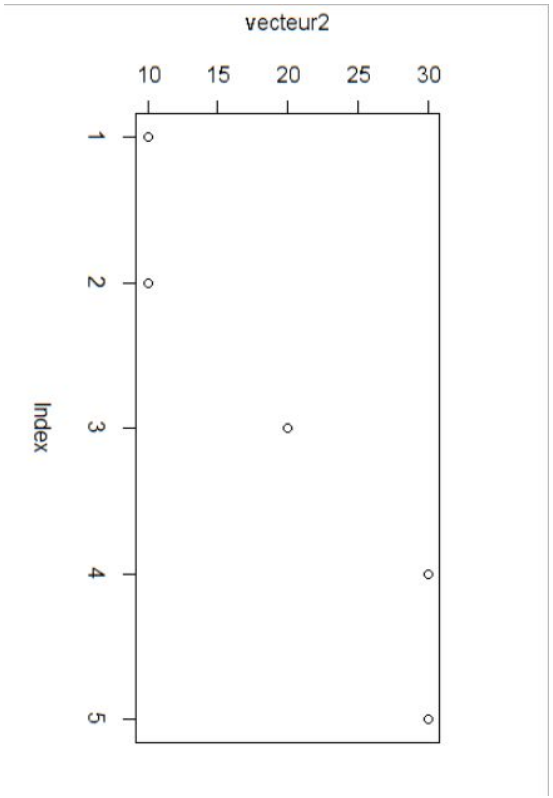
Exemple: vecteur2 <- c(10, 10, 20, 30, 30)

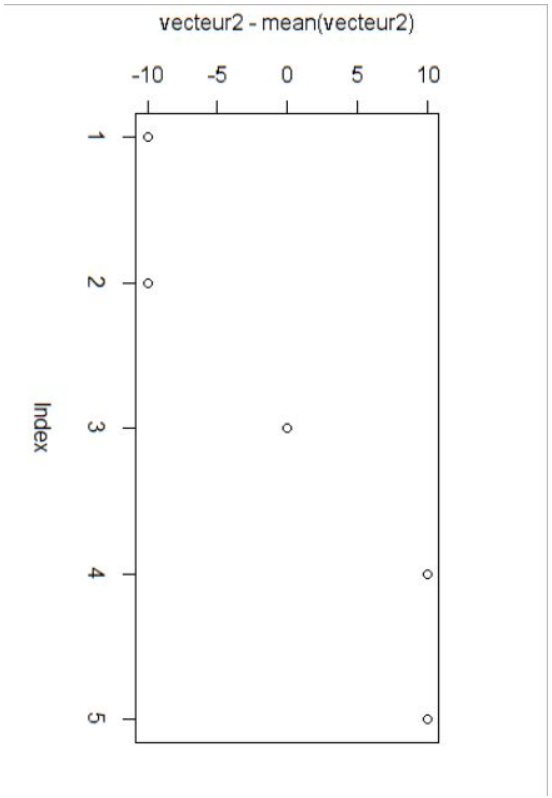
Quel est la variance? (hint: il y a une fonction dans R qui calcule direct)

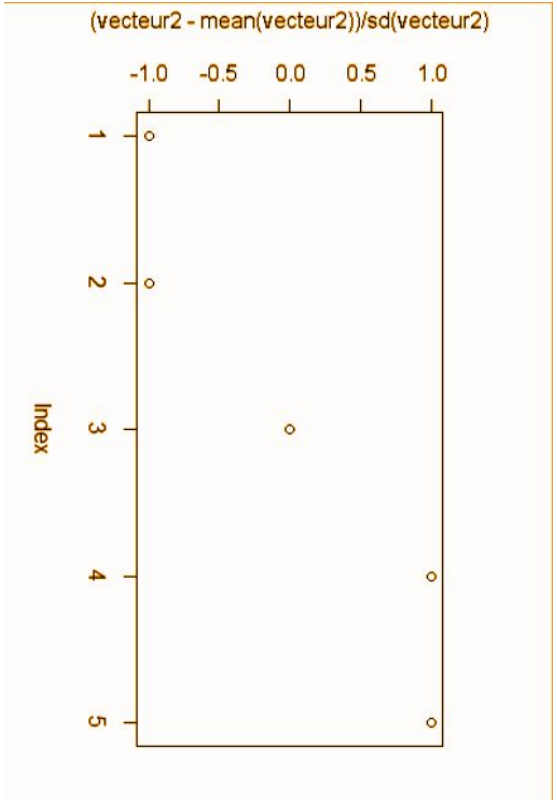
L’ecart type? (pareil)



Z-scores (dernier plot): c’est juste un shift, pas de changement de distribution







Z scores

Transforme des valeurs brutes en z score ne changent pas la forme de la distribution (i.e., on ne va pas obtenir une distribution normale a partir d'une autre etc..). Ca ne fait que bouger la moyenne et la dispersion

Les z score seront utiles plus tard pour:

- Correlation
- Regression
- Null hypothesis significance testing, NHST

Statistiques bivaries

On peut decrire la dispersion/distribution d’une variable avec la variance et l’ecart type

Mais ca peut etre etendu a l’etude de 2 variables:

- Variable X a une certaine distribution
- Variable Y a une certaine distribution
- On peut examiner comment ces deux distributions se comportent ensemble et si elles sont corrélées.

Bivariate dispersion

Imaginons le tableau suivant

height <- c(68, 66, 69, 73, 62, 71)

weight <- c(135, 132, 140, 180, 100, 165)

Comment creer ce tableau dans R a partir des deux vecteurs?

height <dbl>	weight <dbl>
68	135
66	132
69	140
73	180
62	100
71	165

Bivariate dispersion

Chaque variable a sa propre variance mais on peut aussi considérer leur distribution jointe. C’est a dire comment les deux variables bougent ensemble: **covariance.**

On peut facilement visualiser la covariance avec un scatterplot:

ggplot(hwData, aes(height, weight))+

geom_point()

(sinon plot(height, weight) ca marche aussi)

Variance & covariance

Variance X:

$$s^2 = \frac{\sum (X - \overline{X})^2}{N-1} = \frac{\sum (X - \overline{X})(X - \overline{X})}{N-1}$$

Covariance(X, Y)

$$\frac{\Sigma (X - \overline{X})(Y - \overline{Y})}{N-1}$$

Mesure de la covariance

$$\text{cov}_{xy} = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{N - 1}$$

height <dbl>	weight <dbl>	heightDev <dbl>	weightDev <dbl>	crossProduct <dbl>
68	135	-0.1666667	-7	1.166667
66	132	-2.1666667	-10	21.666667
69	140	0.8333333	-2	-1.666667
73	180	4.8333333	38	183.666667
62	100	-6.1666667	-42	259.000000
71	165	2.8333333	23	65.166667

Comment creer ce dataframe avec R?

Quelle est la somme du crossProduct? (i.e. du produit de (X-mean(X))(Y - mean(Y)) .

Quelle est la covariance?

Covariance

La covariance devient:

- Plus positive pour deux variables qui different de leur moyenne dans le meme sens
- Plus negative pour deux variables qui different de leur moyenne dans le sens oppose.

Variance et covariance sont souvent presente dans une matrice variance-covariance: Essaie cov(hwData)

Cela dit, la covariance n'est pas tres pratique pour representer la relation entre deux variables, c'est difficile a interpreter. En particulier son sens change en fonction de l'ecart type:

- Une covariance de 105.8 est:
  - Large si les variables ne sont pas tres dispersees (i.e., petits ecart type)
  - Petite si les variable sont tres dispersees (i.e., grands ecart type)

C'est pour ca qu'on a la **correlation**

Correlation

La correlation = la covariance standardisee, cad on divise la covariance par l'ecart type de chaque variable.

$$\frac{\sum(X-\bar{X})(Y-\bar{Y})}{(SD_x)(SD_y)(N-1)} = \frac{\mathrm{cov}_{xy}}{(SD_x)(SD_y)}$$

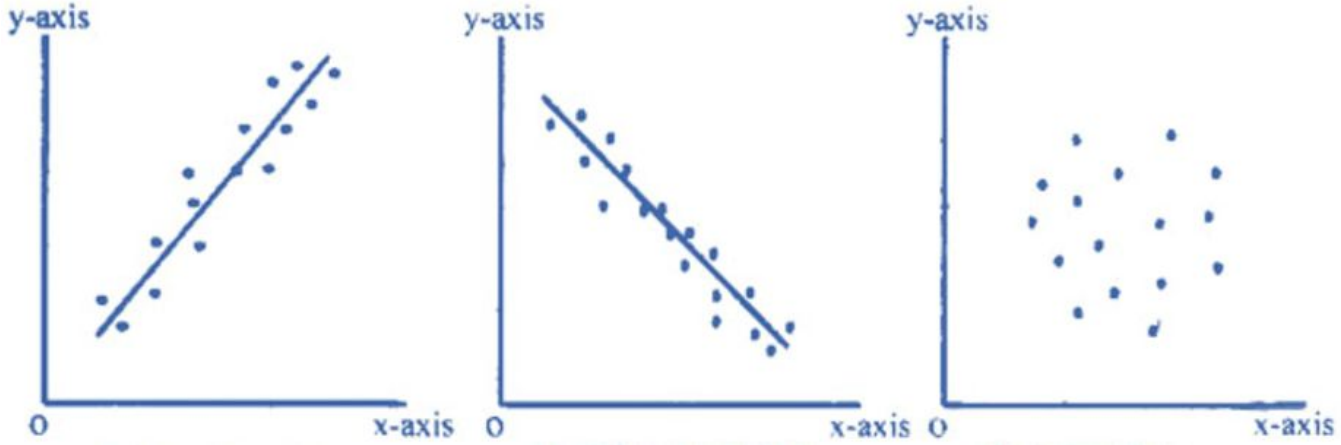
Pearson product-moment correlation (Pearson, 1895)

Pearson product-moment correlation (Pearson, 1895)

Le coefficient de correlation n’a pas d’unite

Il est compris entre -1 et 1

Le signe de la correlation vous dit la direction de la relation entre les deux variables.



Correlation

La valeur du coefficient vous indique si les deux variables sont proches d'un modele lineaire (i.e., une droite).

- r= 1 ou -1 → relation lineaire parfaite, on a une droite
- r = 0 → il n'y a pas de relation lineaire, on a un genre de nuage

En gros, la correlation = est ce que les points sont organise en ligne (ou proche d'une ligne) ou pas.

Correlation

$$\frac{\sum(X-\bar{X})(Y-\bar{Y})}{(SD_x)(SD_y)(N-1)}$$

height <dbl>	weight <dbl>	heightDev <dbl>	weightDev <dbl>	crossProduct <dbl>
68	135	-0.1666667	-7	1.166667
66	132	-2.1666667	-10	21.666667
69	140	0.8333333	-2	-1.666667
73	180	4.8333333	38	183.666667
62	100	-6.1666667	-42	259.000000
71	165	2.8333333	23	65.166667

Quelle est la somme du crossProduct?

Quels sont sd(height) et sd(weight)?

N-1?

Le coefficient de correlation?

Correlation

Avec R bien sur on a pas besoin de faire tout ca parce qu’il y a une fonction dediee: cor()

Essaie cor(height, weight)

Comment peut on interpréter ce nombre?

Correlation - une autre interpretation

$$\frac{\sum(X-\bar{X})(Y-\bar{Y})}{(SD_x)(SD_y)(N-1)}$$

Que representent ces elements?

Correlation - une autre interpretation

$$\frac{\sum(X-\bar{X})(Y-\bar{Y})}{(SD_x)(SD_y)(N-1)}$$

Que representent ces elements? Les z scores pour les variables X et Y. Donc la correlation est la moyenne du produit des z-scores.

Correlation

Comme la covariance, les correlations entre plusieurs variables sont souvent representees sous forme de matrix

Essaie cor(hwData)

C’est quoi les 1 sur la diagonale?

Correlation - interpretation

Cohen a etabli les criteres suivants souvent utiliser dans la recherche:

- r = .1 → faible correlation
- r = .3 → moyenne correlation
- r = .5 → forte correlation

ATTENTION : Correlation n’est pas Causation!!

- Variables confondantes
- Directionalite

Correlation - effect size

L’effect size vous permet d’etablir le percentage de la variance dans vos donnees qui est expliquee par vos variables.

Pour la correlation de Pearson, l’effect size est r^2

Donc si r = .3, r^2 = .09 ⇒ une correlation de .3 signifie que moins de 10% de la variance d’une variable est expliquee par l’autre variable.

Correlation - significance

Il y a plusieurs façons de tester la significance d'une correlation:

- Est ce que le coefficient *r* est different de 0?
- Est ce que deux *r* sont differents l'un de l'autre?
- etc...

Note: *r* est une estimation du parametre de population ρ (rho)

Est ce que *r* differe de 0?

Deux hypotheses:

- H_0: $\rho = 0$ -- cad, X et Y sont indépendantes l'une de l'autre (quasi jamais 100% vrai)
- H_1: $\rho \neq 0$

Pour ce test de significance, on utilise le Student's *t* test:

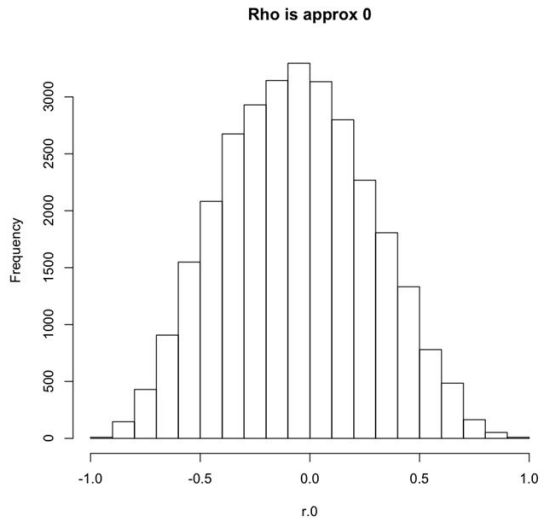
$$T = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

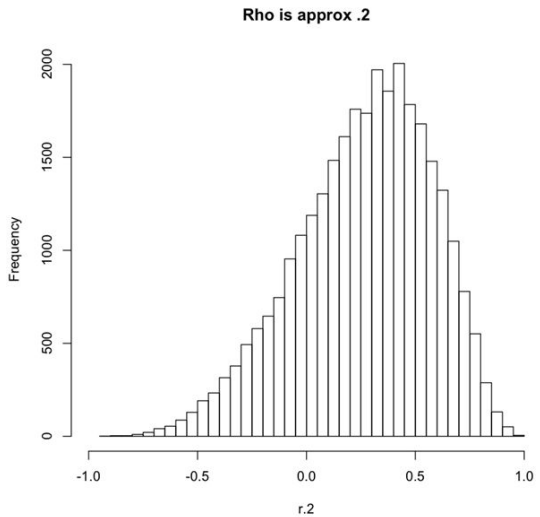
Sur R, on peut obtenir ce test avec la fonction cor.test()

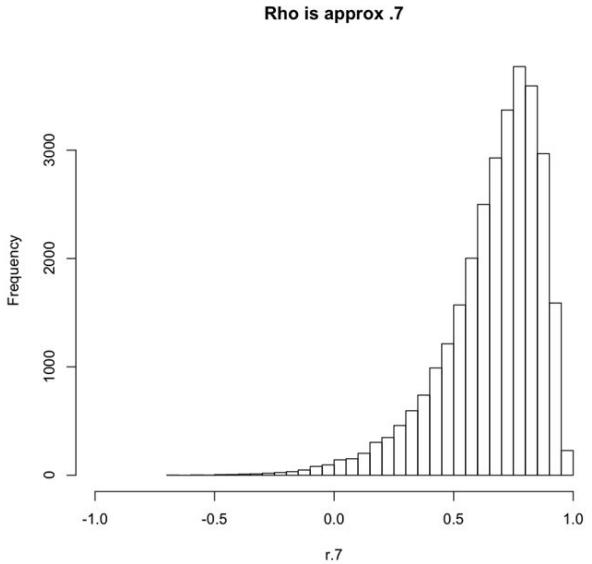
Essaie: cor.test(hwData$height, hwData$weight)

Est ce que deux *r* different l'un de l'autre?

C'est un test un peu plus complique parce que si on fait l'hypothese que $\rho \neq 0$, la distribution d'ou provient *r* n'est pas normale

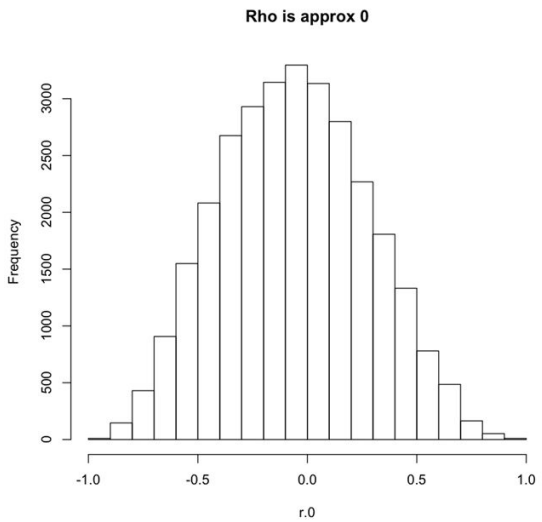


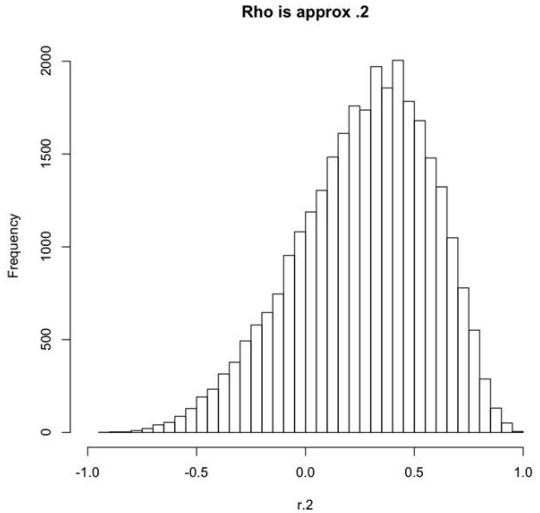


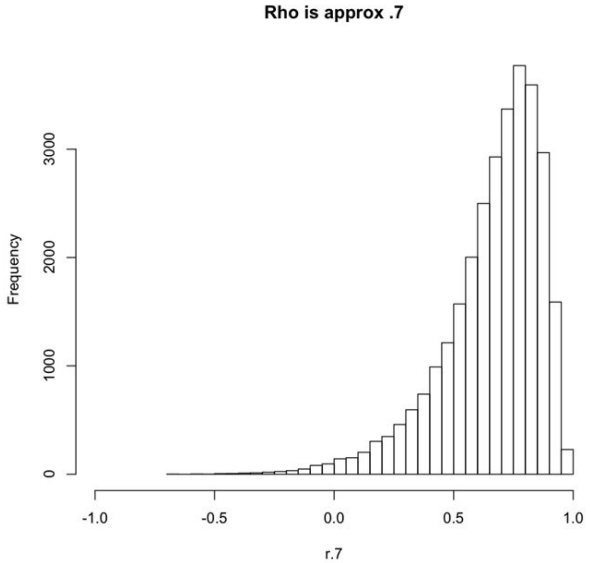


Est ce que deux *r* different l'un de l'autre?

Pour faire un *t* test il faut l'erreur type a la moyenne, qui est l'ecart type de la distribution d'échantillonnage - c'est complique sans la normalite.







Est ce que deux *r* different l'un de l'autre?

Pour eviter ce probleme on utilise Fisher's z

On reviendra sur ce concept plus tard

Correlation - influences

Certaines facteurs peuvent influencer les correlations:

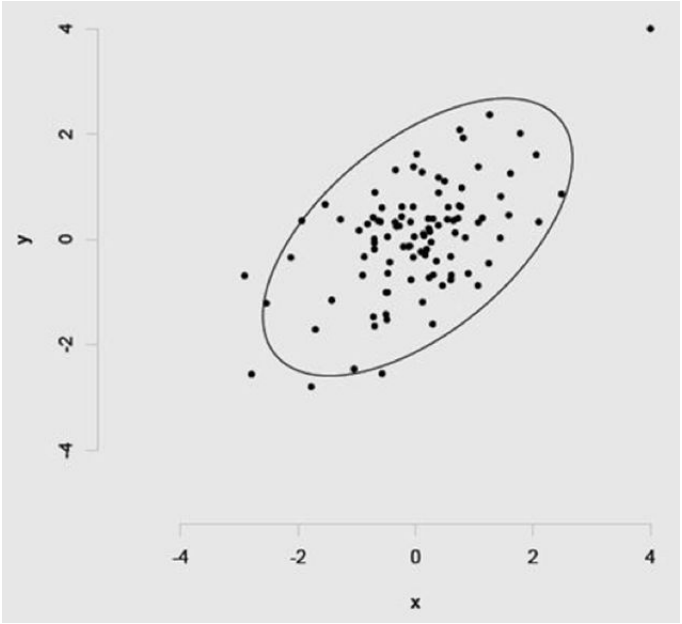
- La presence d’outliers (i.e., de valeurs extremes)
- Probleme de range restriction (i.e., les valeurs que peuvent prendre nos variables)
- Echantillons heterogenes.

Outliers

Ils posent probleme parce qu'ils peuvent generer des correlations qui n'existent en fait pas. Ou les faire paraitre plus forte qu'elles ne sont.

Avec l'outlier: r = .50

Sans l'outlier: r = .57

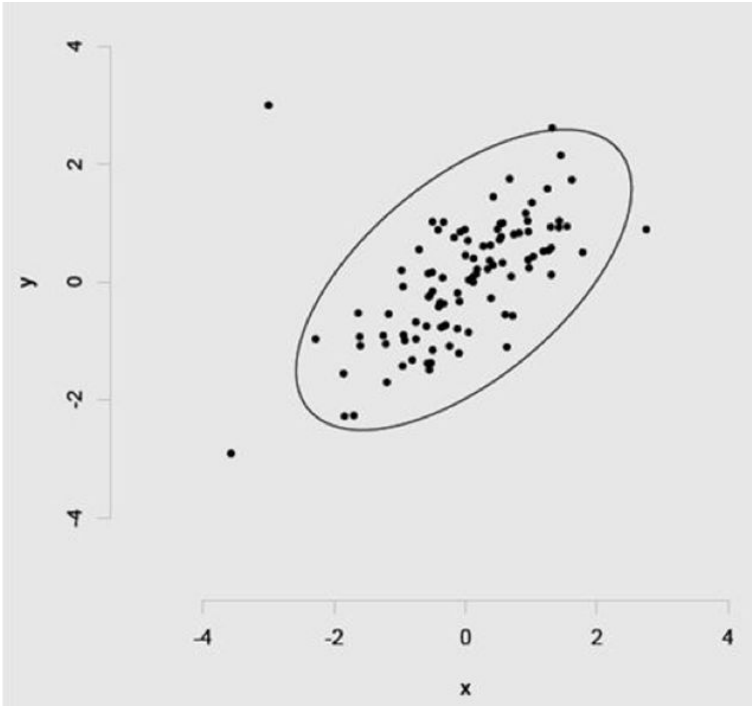


Outliers

A l’inverse ils peuvent atténuer une correlation qui est en fait sans doute assez forte

Avec l’outlier: r = .61

Sans l’outlier: r = .75

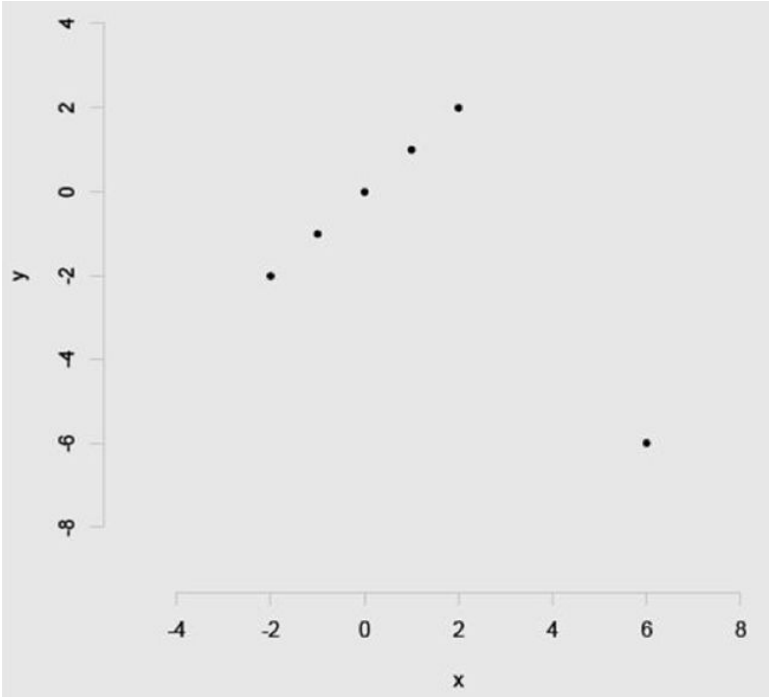


Outliers

Peuvent vraiment poser de gros soucis, allant jusqu'a change le signe du coefficient de correlation

Avec l'outlier: r = - 0.5

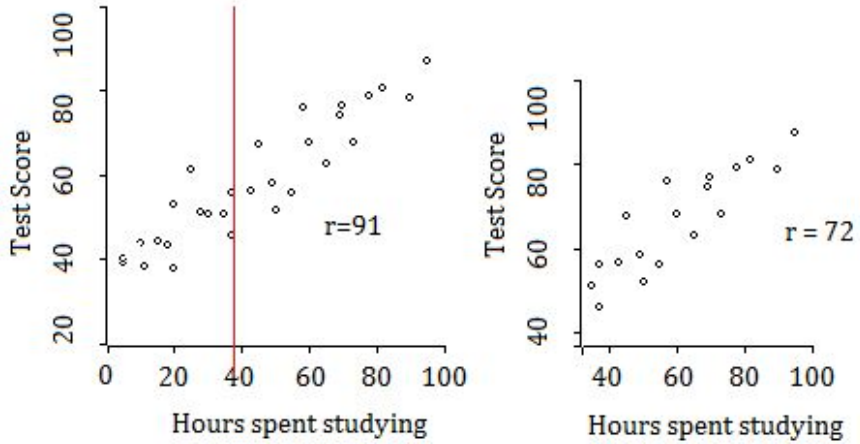
Sans l'outlier: r = 1



Range restriction

C’est a dire que tu ne regardes qu’un subset des valeurs que peuvent prendre tes variables.

Le coefficient *r* aura tendance a baisser sur des valeurs restreintes.

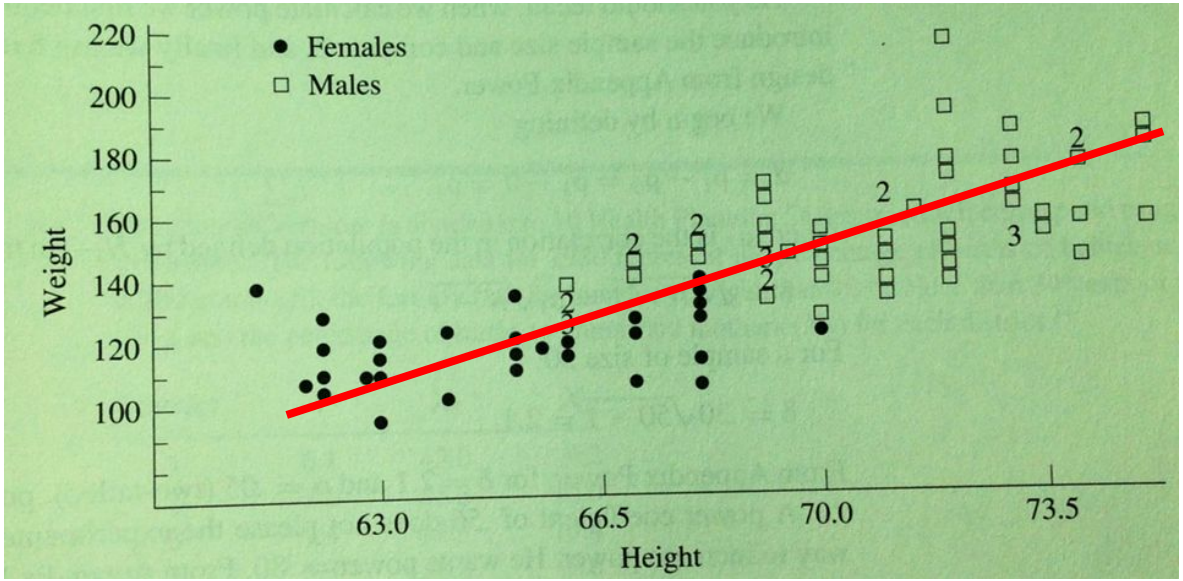


Echantillon heterogene (i.e., extra variable)

Imagine un échantillon de 92 etudiants ou tu mesures le poids et la taille des gens. On obtient ca:

Ca donne un r = .78

Mais...

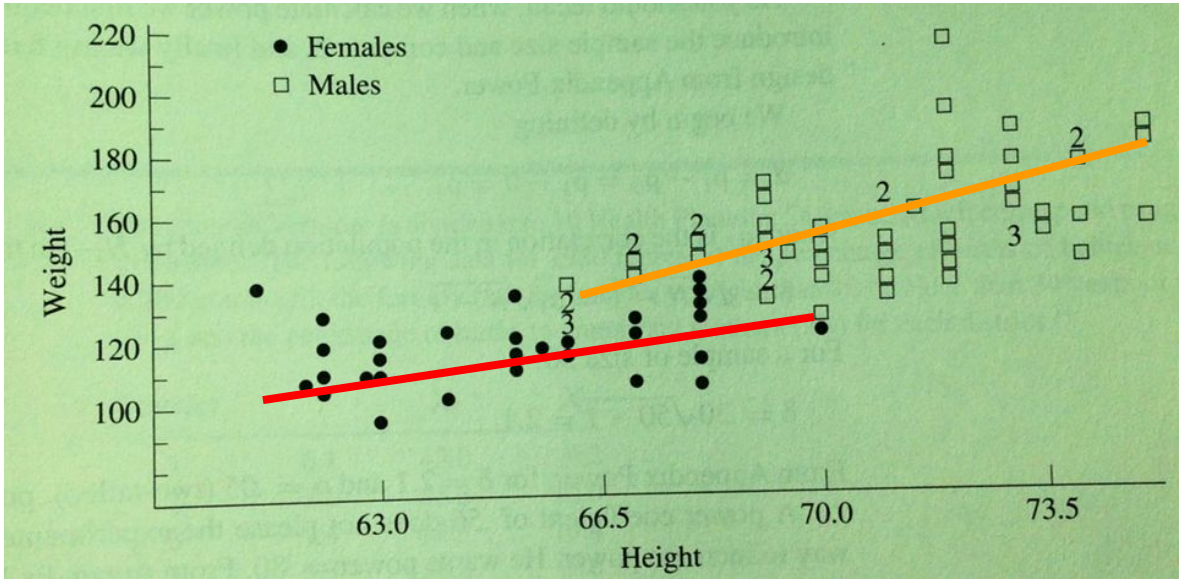


Echantillon heterogene (i.e., extra variable)

Si on considere Males & Females a part on obtient:

rMales = .60

rFemales = . 49



Correlation

Il y a d’autres aspects a discuter:

- Correlation pour des variables qui ne sont pas normalement distribuée
- Semi-correlation
- Correlation power (pouvoir)

On en reparlera plus tard si on a le temps.

Pour la semaine prochaine

- Faire la fiche d’exercice Correlation (deadline Vendredi 26 a 8h30 am)

Semaine pro:

- Data cleaning (identifier les outliers) + plotting
- Plus de discussion sur le NHST, intro a la regression