GC – Probabilités et Statistique

http://moodle.epfl.ch/course/view.php?id=14271

Cours 9

- (Bref ! !) révision : TCL, tests d'hypothèses
- Distribution t de Student, t-test
- Processus de recherche, études
- Modélisaton statistique
- Données bivariées
- Modélisation des données bivariées
- Régression linéaire simple
- Distribution de Y conditionnelle sur X
- Distribution d'échantillonnage des paramètres

Révision : Théorème Central Limite (TCL)

Théorème (TCL) : Soient $X_1, X_2, \ldots$ des variables aléatoires indépendantes et identiquement distribuées (iid), et telles que $E[X_i] = \mu$ et $Var(X_i) = \sigma^2 < \infty$ existent. Alors, la distribution de

$$\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$$

se rapproche d'une distribution normale lorsque $n \to \infty$.

C.-à-d. : Plus n est grand ('suffisament grand'), plus *la loi de la somme (ou la moyenne)* se rapproche d'une distribution normale.

$\Rightarrow$ Donc, $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$; $\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$

Révision : Étapes d'un test d'hypothèses

1. **Identifier** le paramètre de la population
2. **Formuler** les hypothèses NULLE et ALT
3. Calculer la **statistique de test**
4. Calculer la ***p*-valeur** p_{obs}
   - p_{obs} est la probabilité d'obtenir une valeur de T *aussi extrême ou plus* (aussi loin de ce qu'on espère ou mm̂e plus, dans la direction de l'ALT) que ce qu'on a obtenu, *EN SUPPOSANT QUE L'HYPOTHÈSE NULLE EST VRAIE*
5. *Règle de décision et interprétation pratique* : on REJETTE l'hypothèse NULLE H si $p_{obs} \leq \alpha$

À propos des échantillons petits...

- Le z-test qu'on a étudié suppose que la distribution d'échantillonage de la statistique de test T est *Normale*, soit
  - exactement, ou
  - approximativement, selon le TCL
- Pourtant :
  - Si les données sont Normalement distributées, ET
  - si l'écart-type (SD) de la population σ est *inconnu*, ET
  - la taille de l'échantillon est *petite* (par exemple, au-dessous de 30)

  ALORS : la vraie distribution d'échantillonage de T possède des *queues plus lourds* que ceux de la distribution Normale
- Dans ce cas, on utilise le *t-test*

‘Student’ (= William Sealy Gosset)

W. S. Gosset

Guinness

Distribution de T quand σ^2 est inconnue

- Rappelons la statistique de test $T = (\overline{X} - \mu_0)/(\sigma/\sqrt{n})$
- **Si** la taille de l'echantillon n est 'suffisament grande', alors sous H, $T \sim N(0,1)$ *quelle que soit la distribution de X* (TCL)
- **Si** les observations $X_1, \ldots, X_n \sim N(\mu_0, \sigma^2)$, alors $T \sim N(0,1)$ pour σ^2 *connue*, *quelle que soit la taille de l'echantillon n*
- **MAIS :** Si la taille de l'echantillon n est *petite*, et la variance σ^2 est *inconnue*, la *vraie* distribution de T a *davantage de variabilité* que la distribution normale (due à l'estimation *imprécise* de σ basée sur peu d'obs)
- Dans le cas **(1)** $X_1, \ldots, X_n \sim N(\mu_0, \sigma^2)$; **(2)** n est petite ; et **(3)** σ^2 est inconnue, alors $T = \dfrac{\overline{X} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$, la distribution t de Student, avec $n-1$ *degrés de liberté* (df ; 'degrees of freedom')
- (La distribution de T dépend du nombre d'observations n)

Distribution t de Student

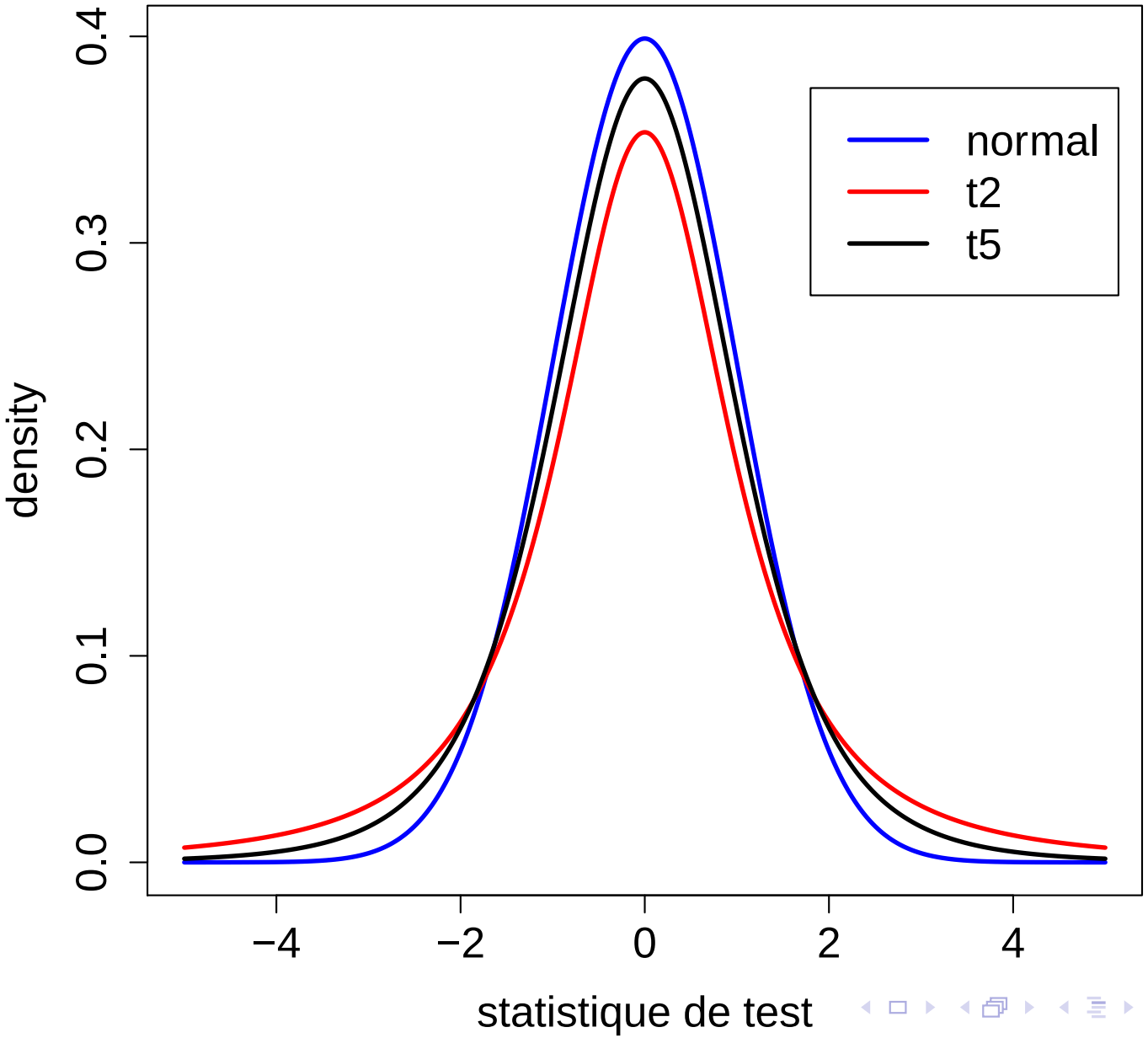


Table de la distribution t de Student

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	**0.50**	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
two-tails	**1.00**	**0.50**	**0.40**	**0.30**	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**0.002**	**0.001**
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

Intervalle de confiance

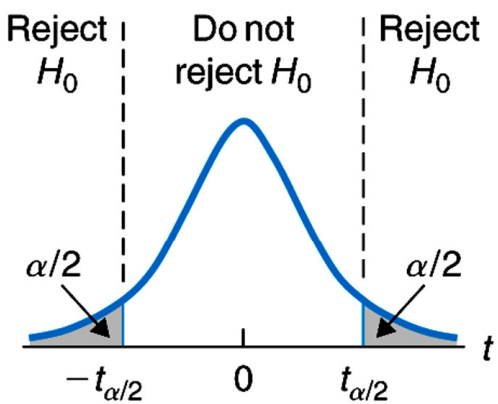
Dans le cas

1. $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$
2. n est petite ; et
3. σ^2 est inconnue :

- on peut faire un *intervalle de confiance (IC)* comme avant, mais **en utilisant la distribution t au lieu de la normale (z)**
- IC pour la *moyenne* de la population : $\overline{x} \pm \boxed{t_{n-1,1-\alpha/2}} \boxed{s} / \sqrt{n}$

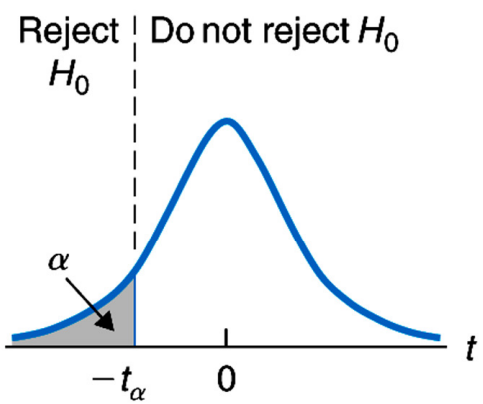
Test d'hypothèses : trouver la région de rejet

H: $\mu = \mu_H$
A: $\mu \neq \mu_H$



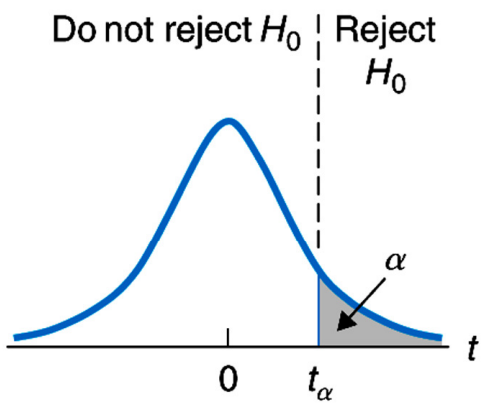
Two-tailed

H: $\mu = \mu_H$
A: $\mu < \mu_H$



Left-tailed

H: $\mu = \mu_H$
A: $\mu > \mu_H$



Right-tailed

Exemple

Exemple 9.1 Prise quotidienne d'énergie (kJ) pour 11 femmes :

5260 5470 5640 6180 6390 6515 6805 7515 7515 8230 8770

- Faire un IC de 95% pour la moyenne prise (kJ) de la population des femmes ...

- Tester l'hypothèse que la moyenne est égale à la valeur recommandée (7725 kJ) ...

Test

1

2

3

4

5

Test de comparaison de 2 moyennes : variances égales

- On veut comparer les moyennes de deux suites de mesures :
  - Groupe 1 (p. ex. 'contrôle') : $x_1, \ldots, x_n$
  - Groupe 2 (p. ex. 'traitement') : $y_1, \ldots, y_m$
- On peut *modeliser* de telles données comme :

  $x_i = \mu + \epsilon_i; i = 1, \ldots, n;$
  $y_j = \mu + \Delta + \tau_i; j = 1, \ldots, m,$

  où Δ signifie l'effet du traitement (par rapport au groupe 'contrôle')
- $H : \Delta = 0$ vs. $A : \Delta \neq 0$ ou $A : \Delta > 0$ ou $A : \Delta < 0$

Variances égales, cont.

- T = diff. observée. / ES(diff. observée.) = $\dfrac{\Delta}{\sqrt{Var(\hat{\Delta})}}$;

  $\hat{\Delta} = \bar{y} - \bar{x}$; $Var(\hat{\Delta}) = \dfrac{\sigma^2}{n} + \dfrac{\sigma^2}{m} = \dfrac{n+m}{nm}\sigma^2$
- On suppose que :
  - les variances des 2 échantillons sont *égales* :
    $Var(\epsilon) = Var(\tau)$
  - les observations sont *indépendantes*
  - *les 2 échantillons sont indépendants*
- On peux estimer les variances *séparement* :
  $s_x^2 = ((x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2)/(n-1)$
  $s_y^2 = ((y_1 - \bar{y})^2 + \cdots + (y_m - \bar{y})^2)/(m-1)$
- Quand les variances sont *égales*, on peut combiner les deux estimateurs : $s_p^2 = ((n-1)s_x^2 + (m-1)s_y^2)/(n+m-2)$

  $$\Rightarrow t_{obs} = \frac{\bar{y} - \bar{x}}{\sqrt{s_p^2(n+m)/(nm)}} \sim t_{n+m-2} \text{ sous } H$$

Test de comparaison de 2 moyennes : variances inégales

- Si $\sigma_x^2 \neq \sigma_y^2$, on peut utiliser

$$T_{Welch} = \frac{\overline{Y} - \overline{X}}{\sqrt{S_x^2/n + S_y^2/m}}$$

- La distribution de cette statistique T_{Welch} *n'est qu'approximativement* t, avec un nombre de degrés de liberté calculé à la base de s_x, s_y, n et m
- Welch test
- Dans la pratique, si les variances sont assez différentes (rapport plus de 3), on utilise cette statistique (au lieu de celle avec la variance s_p^2)

Exemple

Exemple 9.2 Dépenses d'énergie dans les groupes de femmes minces et obèses :

mince	7.53	7.48	8.08	8.09	10.15	8.40	10.88	6.13	7.90	7.05	7.48	7.58	8.11
obese	9.21	11.51	12.79	11.85	9.97	8.79	9.69	9.68	9.19				

- Tester l'hypothèse que les moyennes des deux populations sont égales ...

Test

1

2

3

4

5

Expériences appariées

- Pour une expérience effectuée en *blocs de deux unités*, la *puissance* du t-test pourrait être augmentée
- Cette idée permet *d'éliminer les influences d'autres variables* (p. ex. l'àge, le sexe, *etc.*), en leur donnant des 'traitements' différents
- Ainsi, on a une comparaison des deux conditions *plus précise*

t-test pour une expérience appariée

- Les données sont de forme :

	1	2		n	
contrôle	x_1	x_2	$\cdots$	x_n	espérance μ
traitement	y_1	y_2	$\cdots$	y_n	espérance $\mu + \Delta$

- *Chaque bloc* nous permet d'évaluer l'effet du traitement
- En effet, on considère *les différences*

$$d_1 = y_1 - x_1, \ldots, d_n = y_n - x_n$$

comme un échantillon de mesures provenant d'une distribution d'espérance Δ

- $H : \Delta = 0$ vs. $A : \Delta \neq 0$ ou $A : \Delta > 0$ ou $A : \Delta < 0$
- $T = t$-apparié $= \frac{\overline{d}}{s_d/\sqrt{n}}$, où
  $s_d^2 = ((d_1 - \overline{d})^2 + \cdots + (d_n - \overline{d})^2)/(n-1)$
- Sous H, t-apparié $\sim t_{n-1}$

Exemple 9.1, cont.

Exemple 2.2, cont. : Prise quotidienne d'énergie des 11 femmes pré- et post-menstruel :

pré	5260	5470	5640	6180	6390	6515	6805	7515	7515	8230	8770
post	3910	4220	3885	5160	5645	4680	5265	5975	6790	6900	7335

- Tester l'hypothèse qu'il n'y a pas de différence de prise quotidienne pré et post ...

Test

1

2

3

4

5

Processus de recherche

- *Question* d'intérêt scientifique
- Décider : *quelles données* à recueillir (et comment)
- Collecte et *analyse* des données
- Conclusions, généralisations : *inférence* sur la population
- *Communication* et diffusion des résultats

Question Générique :
Est-ce qu'un 'traitement' produit-il un 'effet' ?

Exemples :

- Fumer provoque-t-il le cancer, les maladies cardiaques, *etc* ?
- Est-ce que la consommation d'avoine diminue le taux de cholestérol ?
- L'échinacée prévient-elle le maladies ?
- Est-ce que l'exercice ralentit le processus de vieillissement ?

Genres d'études

- Une méthode simple pour résoudre ce type de question consiste à *comparer deux groupes* de sujets de l'étude :
  - *Groupe contrôle* : fournit une base de comparaison
  - *Groupe traitement* : groupe recevant le 'traitement'
- *Étude expérimentale* : sujets affectés aux groupes (traitement, contrôle) par l'investigateur
  - *randomisation* : protège contre les biais dans l'attribution aux groupes
  - *'aveugle', 'double-aveugle'* : protège contre les biais dans l'évaluation des résultats
  - *placebo* : traitement artificiel
- *Étude d'observation* : sujets 'attribuent' eux-mêmes aux groupes
  - *facteur de confusion* : un facteur qui présente une association avec le facteur de risque examiné *et* avec le résultat

Quelques commentaires

- Avec une expérience contrôlée bien planifiée et exécutée, il est possible de déduire *la causalité*
- Ceci *n'est pas possible* avec les études d'observation en raison de la présence de facteurs de confusion
- En présence de facteurs de confusion, il n'est pas possible de dire si la différence observée entre les groupes est due au *traitement* ou au *facteur de confusion*
- Pas toujours possible de mener une étude expérimentale, pour des raisons *pratiques* et *éthiques*

Modèles statistiques

- Un **modèle statistique** est une description mathématique approximative du mécanisme qui a généré les observations, qui tient compte des *erreurs aléatoires et imprévisibles* :
  - donne une représentation *idéalisée* de la réalité
  - fait des *suppositions explicites* (qui peuvent être **fausses** ! !) sur les processus étudiés
  - permet un raisonnement *abstrait*
- Le modéle s'exprime par une *famille de distributions théorique* qui contient des cas 'idéaux' pour les VAs inclues
  - p. ex. : jets d'une pièce **...**
- Un modèle utile offert un *bon compromis* entre
  - description *juste* de la réalité (paramètres nombreux, suppositions correctes)
  - *facilité* de manipulation mathématique
  - production de solutions/prédictions *proches* de l'observation(s)

Un modèle simple

Un cas simple : on effectue plusieurs mesures d'une quantité physique μ, p. ex. longueur d'un champ, taille d'une personne ...

- De telles mesures possèdent en général une composante *aléatoire* due aux *erreurs de mesure*
- Un mécanisme d'erreur possible :

$$\begin{array}{ccccc} \text{mesure} & = & \text{vraie valeur théorique} & + & \text{erreur de mesure} \\ y & = & \mu & + & \epsilon \end{array}$$

- c.-à-d. : des mesures avec des *erreurs additives*
- S'il n'y a pas d'erreur *systématique* (biais), l'erreur aléatoire doit être 'centrée' ($E[\epsilon] = 0$)
- Souvent raisonnable de penser que *la précision* de chaque mesure est *la même* ($Var(\epsilon) = \sigma^2$ pour chaque mesure)
- *Une spécification possible* pour la distribution de l'erreur est la loi *normale* $N(0, \sigma^2)$
- **All models are wrong ; some are useful**

Estimation des paramètres inconnus

- Une fois un modèle est choisi, l'interêt se tourne vers l'estimation des inconnus : *les paramètres du modèle*
- On observe des *réalisations* d'une VA dont on connaît la distribution (sauf les valeurs des paramètres)
- Donc, on doit *estimer* les paramètres à l'aide des observations $X_1, \ldots, X_n$
- $\hat{\mu} = \overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$
- $\hat{\sigma}^2 = S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$
- L'estimateur S^2 est *nonbiaisé* pour σ^2, et est *indépendant* de celui de μ $(\overline{X})$

PAUSE

Données bivariées

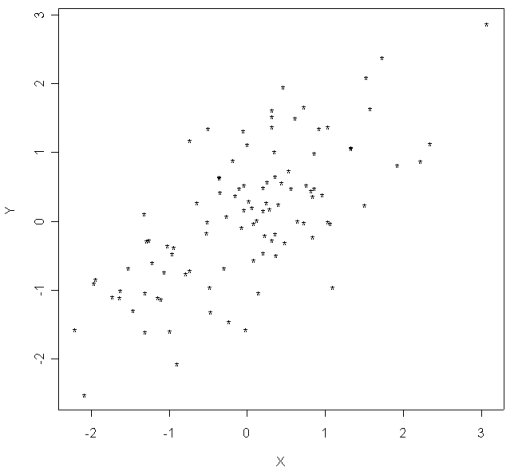
- Mesures de *deux* variables ; p. ex. X et Y
- On considère le cas de deux variables *continues*
- On veut découvrir la *relation* entre les deux variables
  - longueur de l'avant-bras et taille
  - taille et poids
  - expressions de gène A et gène B
- Considérons les ensembles de données qui sont (au moins approximativement)

normales bivariées $\Leftrightarrow$ forme ovale

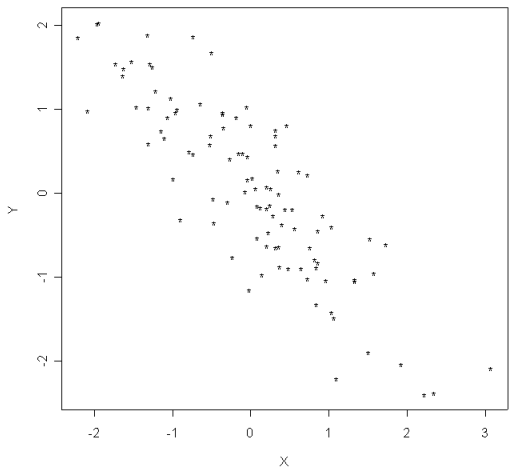
- $(X, Y) \sim BVN((\mu_x, \mu_y), (\sigma_x^2, \sigma_y^2), \rho)$

Analyse exploratoire : Diagramme de dispersion

- **Résumé graphique** d'un jeu de données bivariées à l'aide d'un *diagramme* (ou *nuage*) *de dispersion*
- Les valeurs d'une variable sur l'axe horizontal et les valeurs de l'autre sur l'axe vertical
- Peut être utilisé pour voir comment les valeurs de 2 variables tendent à évoluer les unes avec les autres (c'est-à-dire comment les variables sont **associées**)

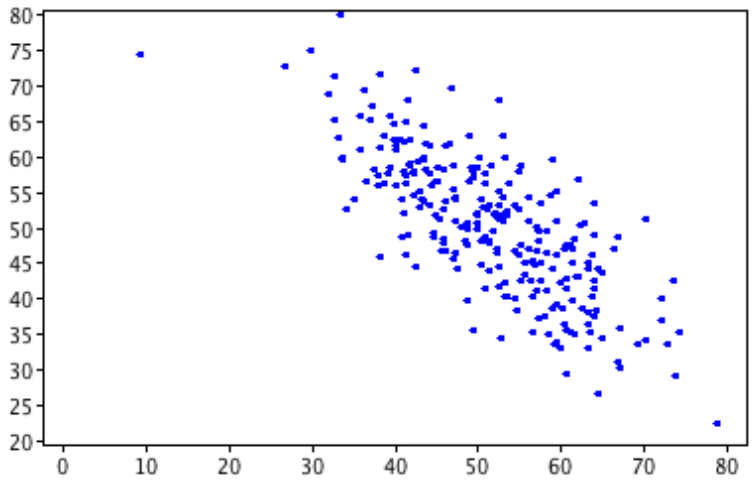


(a) association positive

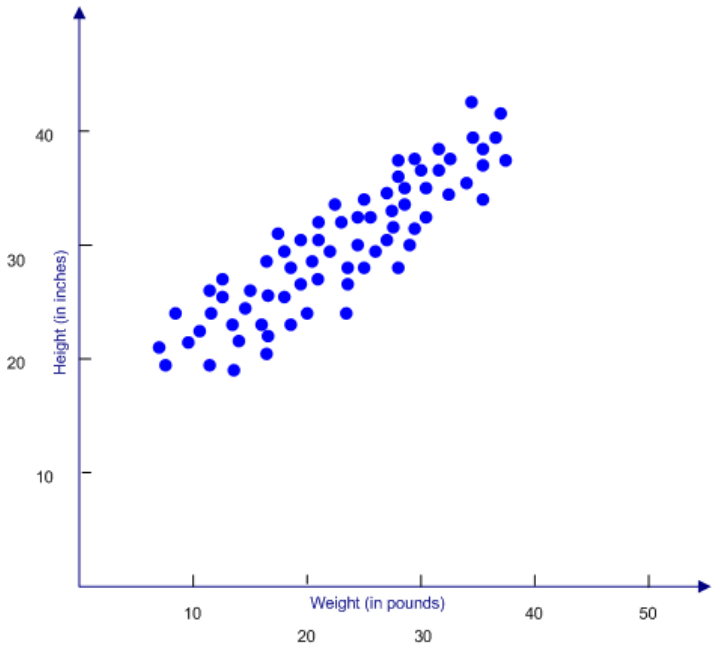


(b) association négative

Diagramme (nuage) de dispersion



(a)



(b)

QCM : *Quelle est l'association entre X et Y ??*

(a) nulle **(b)** positive **(c)** négative **(d)** impossible à déterminer

Figure (a) : _____ Figure (b) : _____

Résumés numeriques

- Typiquement, les données bivariées sont résumées (numériquement) avec **5 statistiques**
- Celles-ci fournissent un bon résumé pour les nuages de points avec la même forme générale que nous venons de voir (ovale)
- On peut résumer chaque variable *séparément* : $\overline{X}, s_X; \overline{Y}, x_Y$
- Mais ces valeurs ne disent pas comment les valeurs de X et Y *varient ensemble*

Corrélation

- Soient X et Y VAs, et $Var(X) > 0, Var(Y) > 0$. La **corrélation** $\rho(X, Y)$ est définie ainsi :

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}} = \frac{E[(X - EX) \times (Y - EY)]}{\sqrt{Var(X)Var(Y)}}$$

- ρ est une *quantité sans unités*, $-1 \leq \rho \leq 1$
- La corrélation ρ, comme la covariance, est **une mesure d'association linéaire** (le degré de linéarité) des VAs X et Y
- Les valeurs ρ proches de 1 ou -1 indiquent une linéarité quasiment rigoueuse entre X et Y, tandis que des valeurs proches de 0 indique une absence de toute relation **linéaire**
- Le signe de ρ indique la direction de l'association (positive ou négative, correspondant à la pente de la droite)
- Lorsque $\rho(X, Y) = 0$, X et Y sont **non-corrélées**

Coefficient de corrélation de l'échantillon

- Le **coefficient de corrélation de l'échantillon** r (ou $\hat{\rho}$) est défini comme la valeur moyenne du produit (normalisé) XY :

$$r = E[(X \text{ centrée-réduite}) * (Y \text{ centrée-réduite})]$$

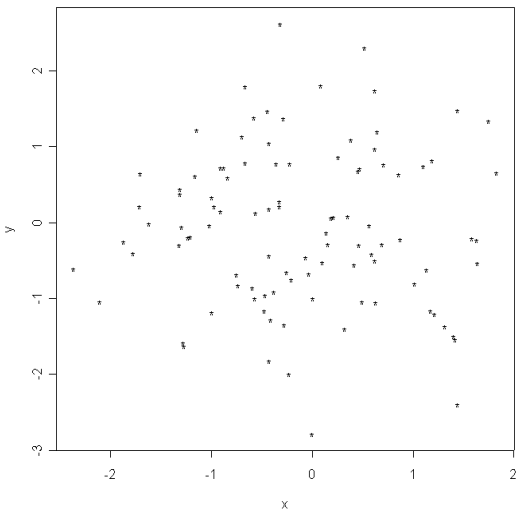
- centrée-réduite = standardisée (normalisée)
  $= (X - \text{ moyenne}(X))/\text{écart-type}(X)$
- r est une quantité *sans unités*
- $-1 \leq r \leq 1$
- r est une mesure d'**ASSOCIATION** **LINÉAIRE**

Corrélation ≠ Causalité

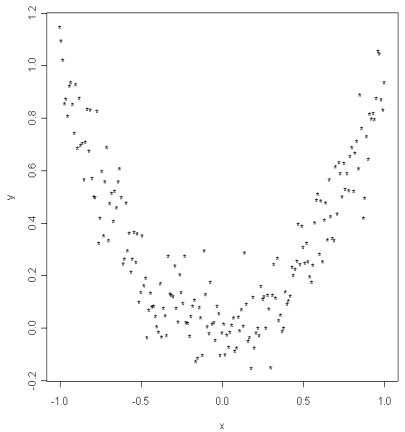
- On *ne peut pas en déduire* que, puisque X et Y sont fortement corrélées (r proche de -1 ou 1) que X est *à l'origine* (ou la *cause*) d'un changement dans Y
- Y pourrait être la cause de X
- X et Y les deux pourraient varier avec un tiers, un facteur peut-être inconnu (soit de causalité ou pas, souvent le temps)
  - polio et boissons non alcoolisées
  - nombre de pompiers envoyés à un incendie et montant des dégâts
  - Les enfants qui reçoivent un soutien scolaire obtiennent de moins bonnes notes que ceux qui ne le reçoivent pas
- Si $r \approx 0$, il n'y a pas d'**ASSOCIATION** **LINÉAIRE**

  – ceci n'est **PAS** à dire qu'il n'y a *AUCUNE ASSOCIATION*
- On ne peut pas en déduire la forme du diagramme de dispersion seulement à partir de la valeur de r

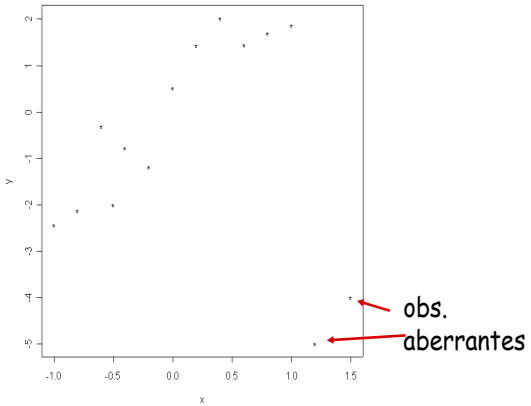
$r \approx 0$



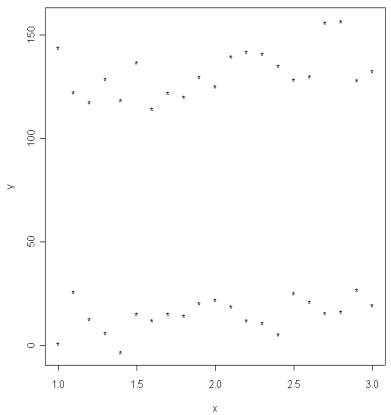
(a) dispersion au hasard



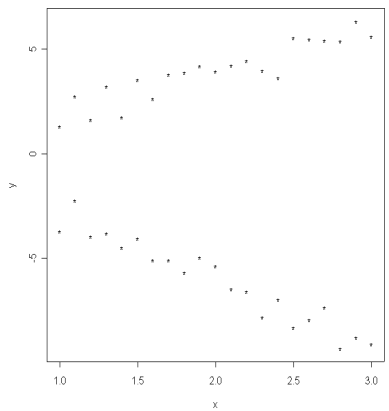
(b) courbe



(c) observations aberrantes



(d) parallélisme

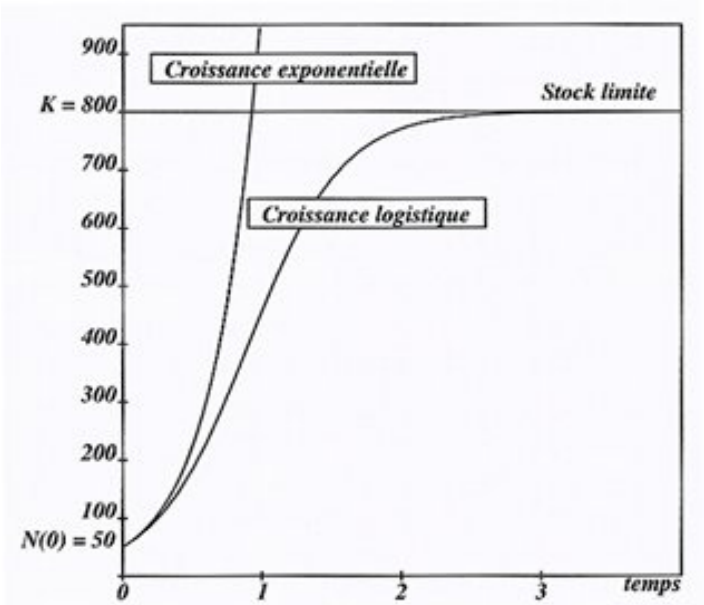


(e) deux droites différentes

Modélisation d'un nuage de forme ovale

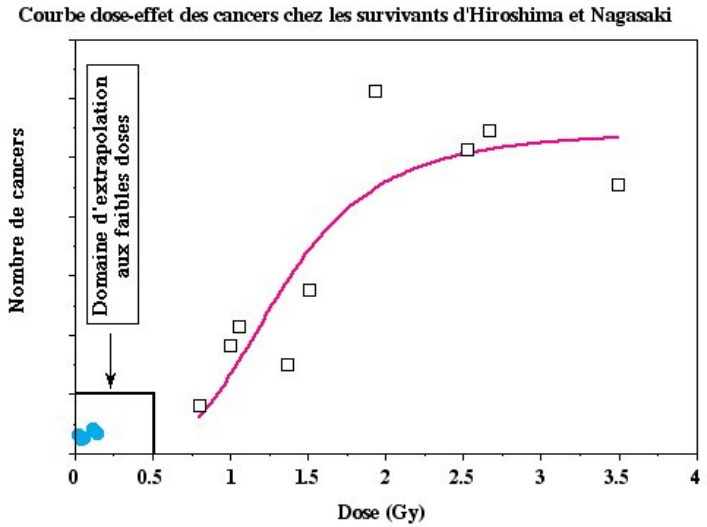
- Variable à expliquer / variable réponse : Y
- Variable explicatrice / prédictrice : X
  - La valeur de X est supposée connue *sans erreur*
  - On suppose que les variations de Y sont *influencées* par X
  - Le modèle permet d'exprimer sous la forme d'une *relation mathématique* la liaison supposée
- La connaissance de ces variables permettent à l'aide du modèle de *prédire* Y
  - Estimater les valeurs de Y :
    - *ponctuellement*
    - par *intervalle*
- Le modèle permet de mesurer *l'impact* (ou *l'effet)* d'une variable explicative sur Y

Relation déterministe ou statistique



(a) déterministe

- *Une seule valeur* de Y pour une valeur de X



(b) statistique

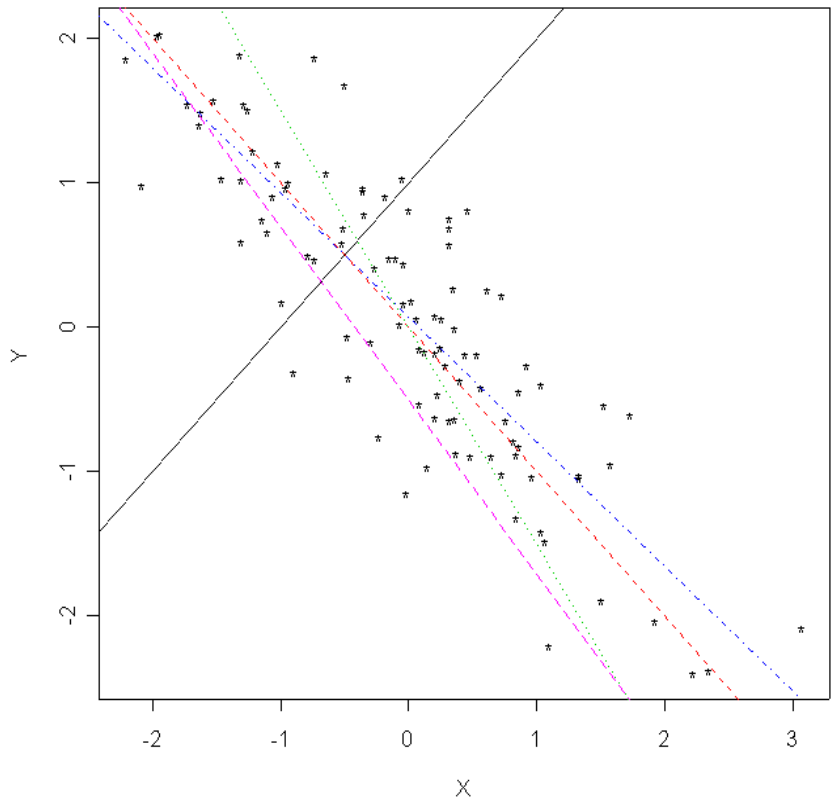
- *Plusieurs valeurs* de Y pour une valeur de X
- *'Probabiliser' Y* pour une valeur fixe de X

Régression linéaire simple

- Se réfère à tracer une droite (particulière) à travers un nuage de points
- Utilisé pour les 2 objectifs :
  - Explication
  - Prédiction
- Modèle linéaire statistique :
  - $Y = \beta_0 + \beta_1 X + \epsilon \quad \Rightarrow \quad E[Y \mid X] = \beta_0 + \beta_1 X$
  - $E(\epsilon) = 0; \;\; Var(\epsilon) = \sigma^2$
- L'équation d'une droite de prédire Y quand on connaît la valeur spécifique x : $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$
- $\beta_0 =$ l'*ordonnée à l'origine*; $\beta_1 =$ la *pente* (dans la population)

Quelle droite ?

- Il y a beaucoup de droites qui pourraient être faites à travers le nuage de points
- Comment choisir ?



Prédiction par régression

- On peut faire une prédiction en utilisant *la droite de régression* :

  lorsque X augmente de 1 (écart-type), la valeur prédite Y augmente ** PAS de 1 (écart-type) **,
  mais seulement de r (écart-type) (vers le bas si r est négatif) :

  - $\dfrac{\hat{Y} - \overline{Y}}{s_Y} = \boldsymbol{r} \dfrac{X - \overline{X}}{s_X}$

- Cette prédiction pourrait s'exprime également dans la forme :

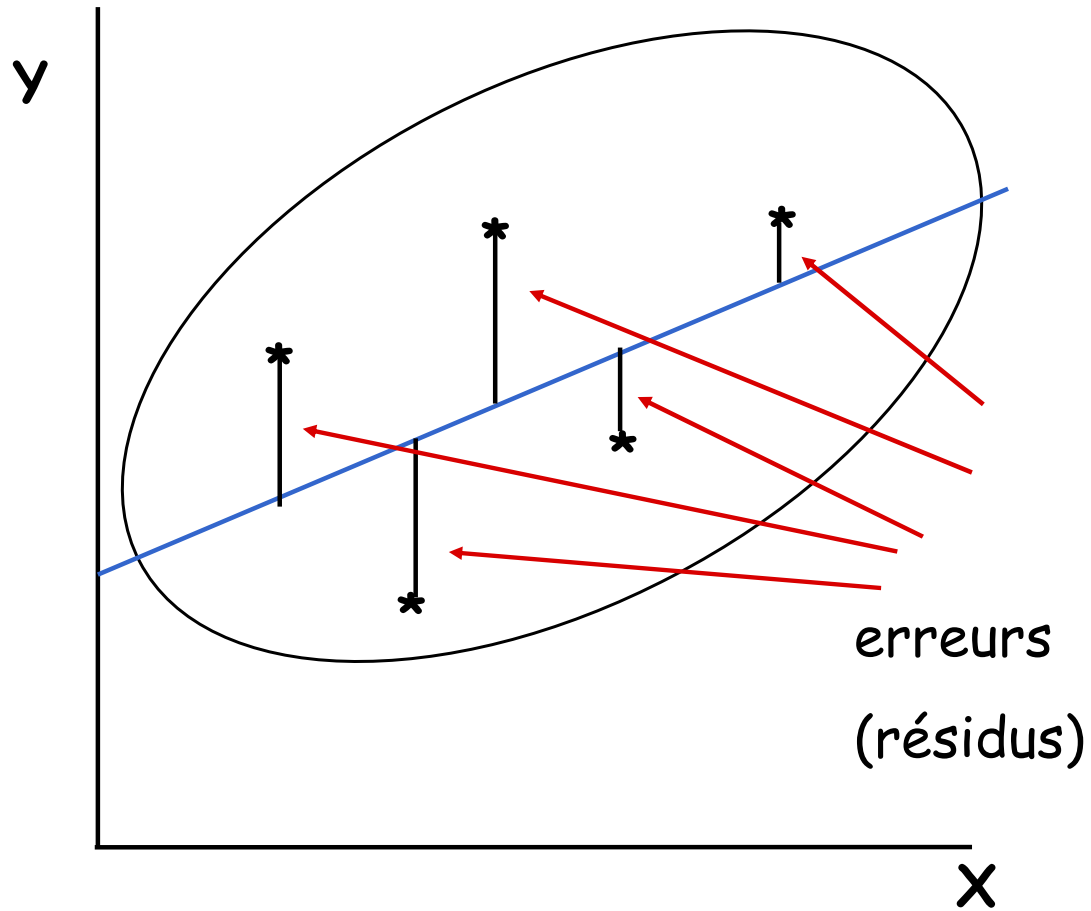
  $$\text{préd. } y = \text{ord.} + \text{pente} \times x, \text{ avec}$$

  - pente $= \hat{\beta}_1 = r\, s_Y / s_X$
  - ord. $= \hat{\beta}_0 = \overline{y} -$ pente $\times \overline{x}$

Moindres carrés

Q : D'où vient cette équation ?

R : C'est la droite qui est 'meilleure' dans le sens que la somme des *carrés des erreurs* dans le plan vertical (Y) est au *minimum*



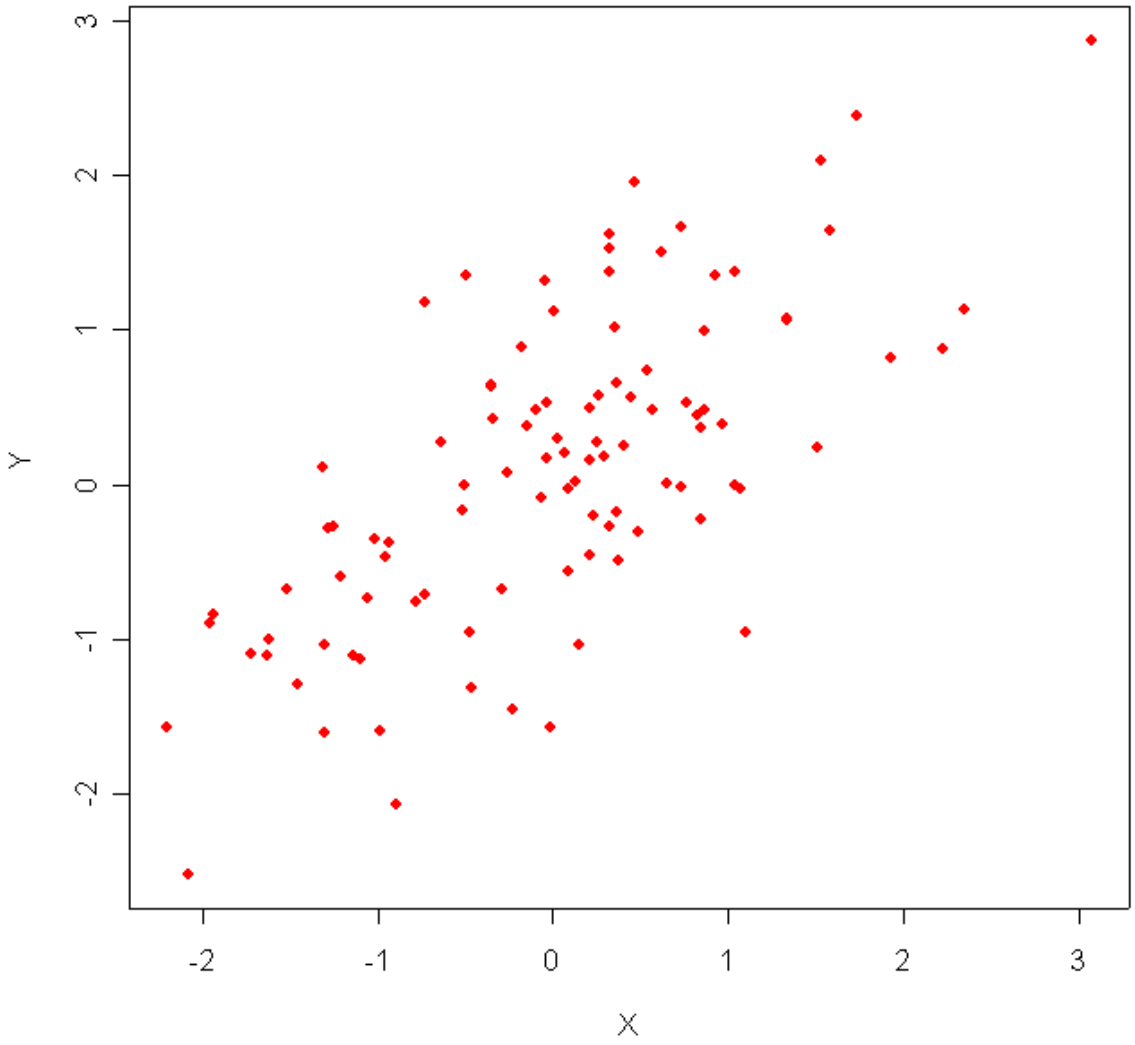
*** Interprétation des paramètres ***

- L'équation de droite de régression comprend 2 paramètres : la *pente* et l'*ordonnée à l'origine*
- La *pente* est le changement moyen de Y pour un changement de X de 1 unité
- L'*ordonnée à l'origine* est la valeur de Y estimée lorsque $X = 0$
- Si la pente $= 0$, alors X n'aide pas à prédire Y (prédiction linéaire)

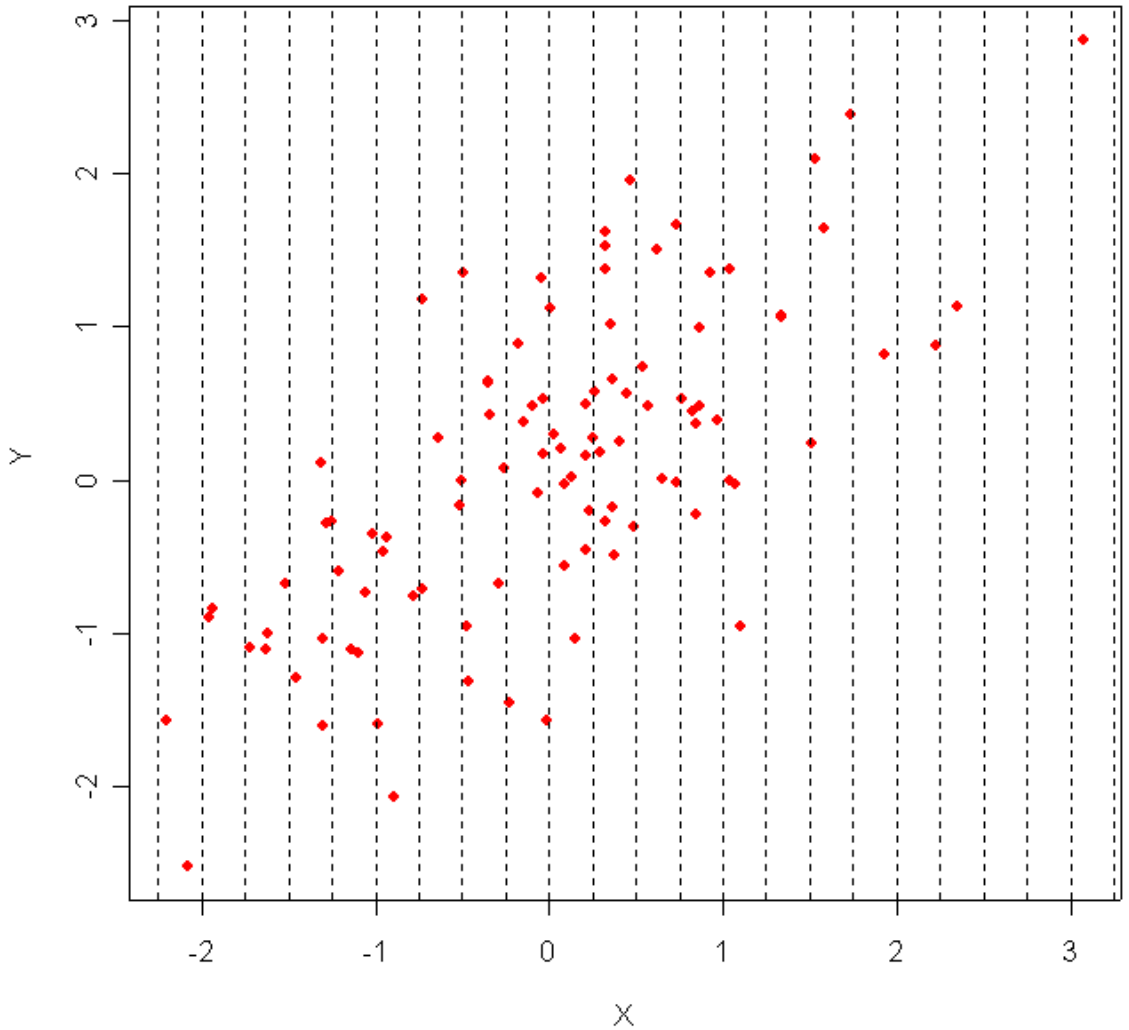
Une autre vue de la droite de régression

- On peut diviser le nuage de points dans les régions (*X-bandes*) fondées sur des valeurs de X
- Au sein de chaque X-bande, mettez la valeur moyenne de Y (en utilisant uniquement les valeurs de Y possèdant des valeurs X dans le X-bande)
- Il s'agit de la *courbe des moyennes*
- La droite de régression pourrait être considérée comme une *version lissée* de la courbe des moyennes

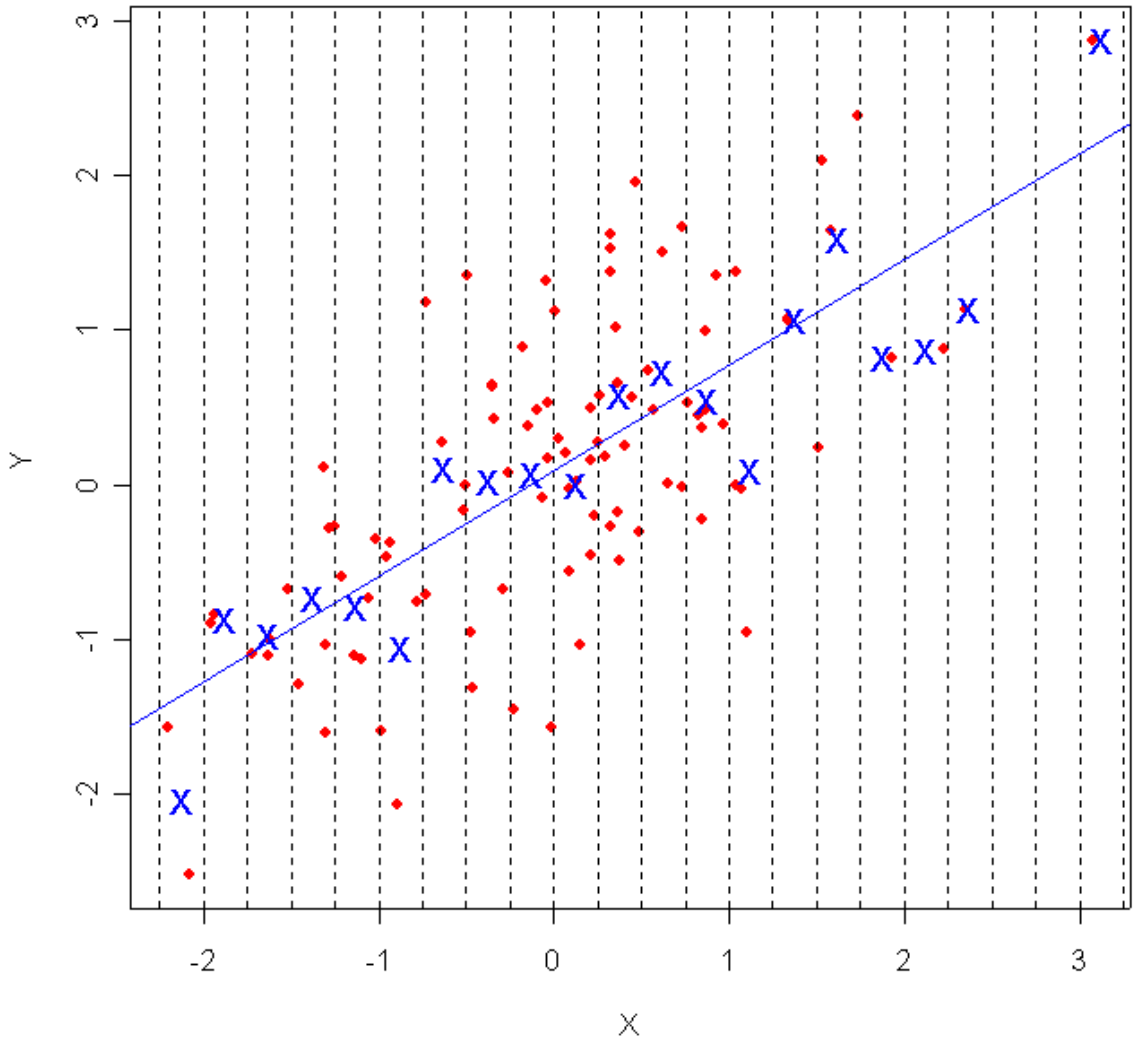
Diagramme de dispersion (encore une fois)



Création des X-bandes



Graphique des moyennes



Démarche de la régression

A partir d'un échantillon de valeurs pour la variable réponse Y et la (ou les) variables prédictrices X :

- Vérifier la possibilité d'une liaison linéaire entre Y et X
  - représentation graphique
  - coefficient de correlation
- Estimation des paramètres
  - coefficients $\beta_i \Rightarrow \hat{\beta}_i$
  - écart-type pour les erreurs $\sigma \Rightarrow \hat{\sigma}$
- Evaluation du modèle (la semaine prochaine)
  - indices de qualité R^2, R^2_{aj}
  - évaluation globale de l'ajustement (F de Fisher)
  - test(s) de coefficients individuellement
  - étude des résidus, détection des points abérrants, influentiels

Résumé : Régression linéaire simple (conceptuelle)

- Pour un diagramme de dispersion qui est *de forme ovale*, nous pouvons trouver une droite qui sert à résumer les points
- Un principe souvent utilisé pour l'ajustement de cette droite est *moindres carrés* : le total des carrés des erreurs (verticales) est réduit au minimum
- Selon ce principe, la prédiction de régression pour Y sachant X nous dit que :
  lorsque X augmente de 1 fois l'écart-type, Y (en espérance) augment de r fois l'écart-type
- On peut trouver l'équation de la droite des moindres carrés en utilisant les 5 statistiques :

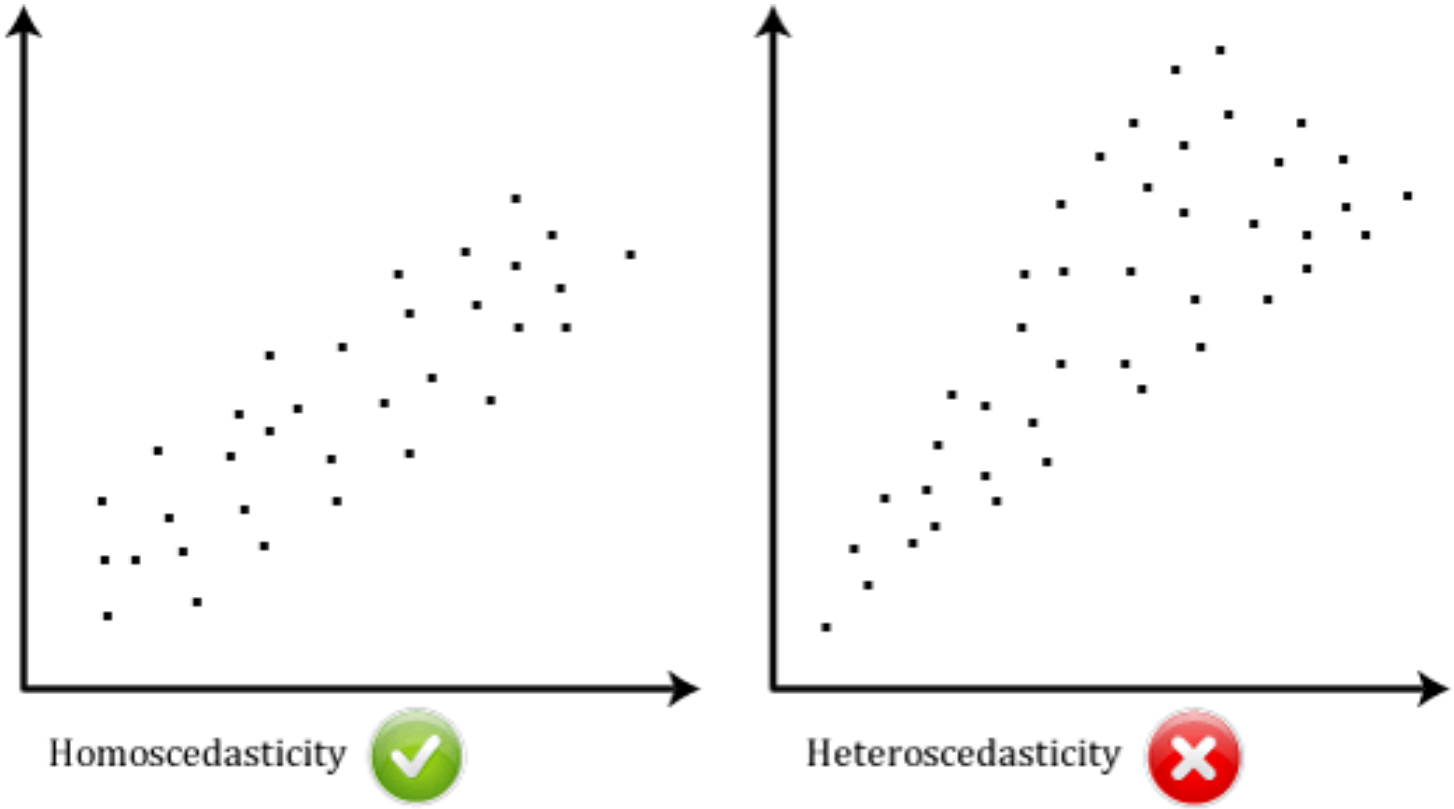
$$\overline{X}, SD(X), \overline{Y}, SD(Y), r$$

- La *pente* (estimée) égale à $\hat{\beta}_1 = r\dfrac{s_Y}{s_X}$,
  l'*ordonnée à l'origine* (estimée) est $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$

Régression linéaire simple – cadre mathématique

- Ici, on considère un modèle où la *variable expliquée* (ou *réponse*) y_i a une association linéaire à une *variable explicative* (ou *régresseur* ou *prédictrice*) x_i :
  $y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n,$
- $\epsilon_1, \ldots, \epsilon_n$ sont supposés variables aléatoires : non corrélées ; espérance $= 0$; variance $= \sigma^2$, $i = 1, \ldots, n$ (*homoscédastique*)
- x_i sont supposés être des constantes (mesurés sans erreur)
- $\Rightarrow$ Si les erreurs sont aussi supposées *normalement distribuées*, on peut faire les *tests* et les *intervalles de confiance (IC)*
- Resumé des suppositions :
  - **L**inear model (modèle linéaire ; dans les paramètres)
  - **I**ndependent errors / observations (erreurs / observations indépendantes)
  - **N**ormal errors / observations (erreurs / observations Normalement distribuées)
  - **E**qual error variances (variances des erreurs égales)

Erreurs homoscédastiques, heteroscédastiques



Méthode des moindres carrés (Les détailles NE SERONT PAS EXAMINÉES)

- Les données ne sont qu'un *échantillon* (et ne sont pas l'ensemble de la population)
- Donc il faut *estimer* les valeurs des paramètres β_0 (ordonnée à l'origine) et β_1 (pente) (également la variance des erreurs σ^2) :

$$\hat{y}_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- Selon le *principe des moindres carrés*, on cherche les estimateurs qui réduisent au minimum :

$$SC(\hat{y}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

- ('SC' = 'somme des carrés' = 'sum of squares' en anglais)

Méthode de moindres carrés, cont.

C'est maintenant *un problème d'optimisation*, de trouver des valeurs $\hat{\beta}_0$ et $\hat{\beta}_1$ qui réduisent au minimum

$$SC(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

Pour résoudre ceci, dériver par rapport à β_0, β_1 ; trouver les zéros :

$$\begin{aligned}
\frac{d}{d\beta_0} &= \sum_{i=1}^{n} -2(y_i - \beta_0 - \beta_1 x_i) = 0 \\
&\Rightarrow \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \\
&\Rightarrow \sum_{i=1}^{n} y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_i = 0 \\
&\Rightarrow \sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i \qquad (*)
\end{aligned}$$

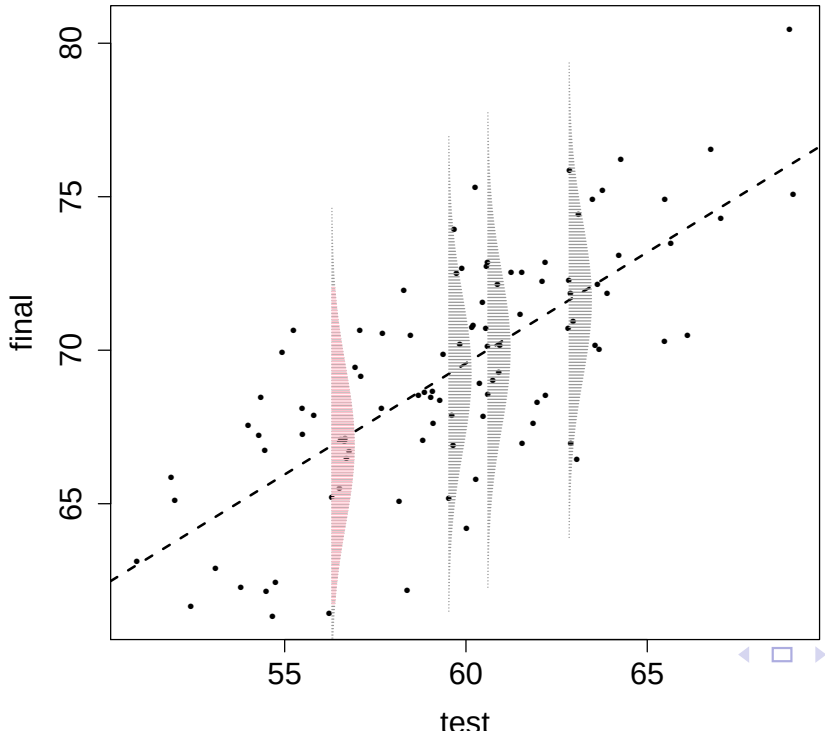
Moindres carrés, cont.

$$
\begin{aligned}
\frac{d}{d\beta_1} &= \sum_{i=1}^{n} -2x_i(y_i - \beta_0 - \beta_1 x_i) = 0 \\
&\Rightarrow \sum_{i=1}^{n} (x_i y_i - \beta_0 x_i - \beta_1 x_i^2) = 0 \\
&\Rightarrow \sum_{i=1}^{n} x_i y_i - \beta_0 \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0 \\
&\Rightarrow \sum_{i=1}^{n} x_i y_i = \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 \qquad (**)
\end{aligned}
$$

Solution simultanée de (*) et de (**) pour les paramètres β_0 et β_1 nous donne **l'estimation de régression**.

Distribution normale conditionelle : graphiquement

- L'*espérance* est la prédiction $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
- L'*erreur-type* est la racine carrée de l'erreur quadratique moyenne :
  EQM = la moyenne arithmétique des carrés des écarts entre les prédictions et les observations
- $REQM(Y) = s_Y \sqrt{(1 - r^2)}$



Distribution normale conditionelle : algèbre

- La *distribution Normale* (univariée) depende de 2 paramtres : la moyenne et la variance (equivalent, le SD)
- L'*espérance conditionnelle* est la prédiction de regression de Y sachant X : $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
- L'*erreur conditionnelle (RMSE)* (racine carré de l'erreur quadratique moyenne EQM) est une nouvelle mesure de variabilité : la variabilité de l'espérance conditionnelle de Y sachant X, *i.e.*, la variabilité autour de la droite de regression ; c'esst la racine carré de *l'erreur quadratique moyenne EQM*
- $MSE(EQM)$ = moyenne arithmétique* des carrés des déviations entre les prédictions et les observations
- *(au lieu de diviser par n, on divise par le nombre de *degrees of freedom* (degrés de liberté))
- $RMSE(Y) = s_Y\sqrt{(1 - r^2)}$

Propriétés de l'estimateur de la pente

- L'estimation de la droite de régression : $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$
- L'estimateur des moindres carrés pour la pente β_1 pourrait être écrit comme :

$$\hat{\beta}_1 = \frac{y_1 (x_1 - \overline{x}) + \cdots + y_n (x_n - \overline{x})}{(x_1 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}$$

- L'espérance de l'estimateur : $E[\hat{\beta}_1] = \beta_1$
- La variance de l'estimateur :

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{(x_1 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}$$

- Il nous faut un estimateur de σ^2 $(e_i = y_i - \hat{y}_i)$:

$$\hat{\sigma}^2 = \frac{e_1^2 + \cdots + e_n^2}{n - 2}$$

Test/Intervalle de confiance pour la pente

- Pour tester $H : \beta_1 = \beta_1^H$ contre $A : \beta_1 \neq \beta_1^H$:

$$t\text{-pente}_{obs} = \frac{\hat{\beta}_1 - \beta_1^H}{\hat{\sigma}/\sqrt{(x_1 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}}$$

- On REJETTE H si : $|t\text{-pente}_{obs}| > t_{\underline{n-2},1-\alpha/2}$
- Le IC de niveau $1-\alpha$ pour la pente β_1 est :

$$\hat{\beta}_1 \pm \frac{\hat{\sigma}}{\sqrt{(x_1 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}} t_{\underline{n-2},1-\alpha/2}$$