GC – Probabilités et Statistique

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

Lecture 3

- (Brief ! !) review : CLT, CI, hypothesis tests
- Student's t distribution, t-test
- Research process, scientific investigations
- Statistical modeling
- Bivariate data
- Modeling bivariate data
- Simple linear regression
- Distribution of Y conditional on X
- Sampling distribution of parameter estimates

Review : Central Limit Theorem (CLT)

- The **Central Limit Theorem** is one of the most important results in probability/statistics, and is widely used as a problem-solving tool
- **Theorem (CLT)** : Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed (iid) RVs, each having mean μ and variance σ^2
- Then for n 'sufficiently large', the distribution of
  - the **sum :** $\sum_{i=1}^{n} X_i$ is approximately $N(n\mu, n\sigma^2)$
  - the **mean :** $\overline{X}$ is approximately $N(\mu, \sigma^2/n)$

Review : steps in hypothesis testing

1. **Identify** the population parameter being tested
2. **Formulate** the NULL and ALT hypotheses
3. Compute the **test statistique (TS)**
4. Compute the *p*-**value** p_{obs}
   - p_{obs} is the probability of obtaining a value of T *as or more extreme* (as far away from what we expected or even farther, in the direction of the ALT) than the one we got, *ASSUMING THE NULL IS TRUE*
5. *Decision rule and practical interpretation* : REJECT the NULL hypothesis H if $p_{obs} \leq \alpha$

Regarding small samples...

- The z-test that we have studied assumes that the sampling distribution of the test statistic T is *Normal*
  - exactly, or
  - approximately, by the CLT
- However :
  - If the data are Normally distributed, AND
  - if the population SD σ is *unknown*, AND
  - the sample size is *small* (for example, under 30)

  THEN : the true sampling distribution of T has *heavier tails* than the Normal distribution
- In this case, you should use the *t-test*

‘Student’ (= William Sealy Gosset)

W. S. Gosset

Guinness

Distribution of T when σ^2 is unknown

- Recall the test statistic $T = (\overline{X} - \mu_0)/(\sigma/\sqrt{n})$
- **If** the sample size n is 'sufficiently large', then under H, $T \sim N(0,1)$ *regardless of the distribution of X* (CLT)
- **If** the observations $X_1, \ldots, X_n \sim N(\mu_0, \sigma^2)$, then $T \sim N(0,1)$ for *known* σ^2, *regardless of the sample size n*
- **BUT :** If the sample size n is *small*, and the variance σ^2 is *unknown*, the *true* distribution of T has *more variability* than the Normal distribution (due to the *imprecise* estimation of σ based on few obs)
- For the case **(1)** $X_1, \ldots, X_n \sim N(\mu_0, \sigma^2)$; **(2)** n small ; and **(3)** σ^2 is unknown, then $T = \dfrac{\overline{X} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$, the Student t distribution, with $n-1$ *degrees of freedom* (df)
- The distribution de T depends on the number of observations n)

Student t distribution

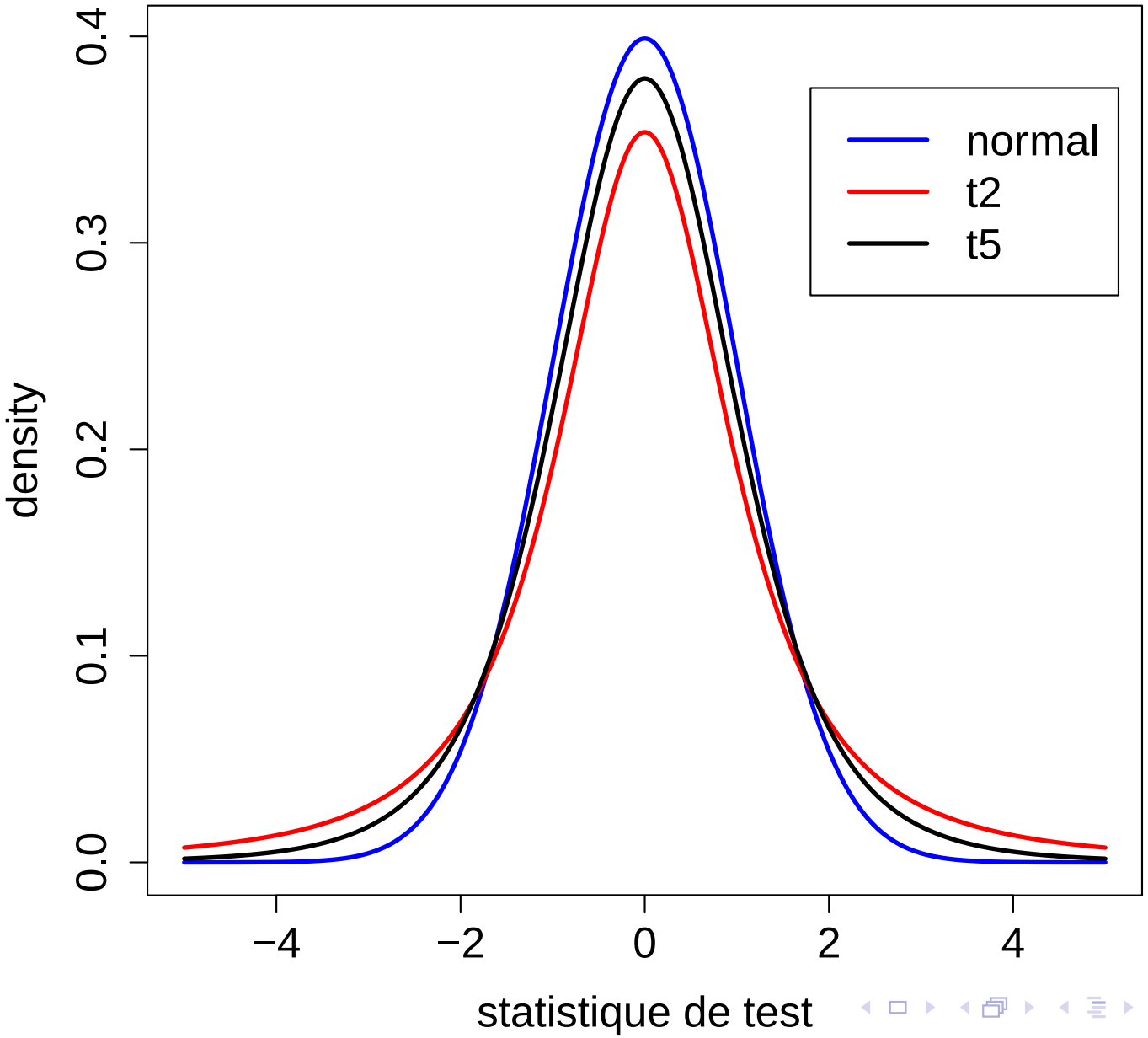


Table of the t distribution

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										

Confidence interval

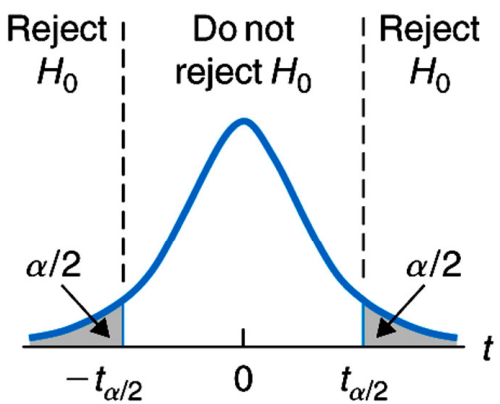
In the case

1. $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$
2. n small ; and
3. σ^2 is unknown :

- we can make a *confidence interval (CI)* as before, but **using the t distribution instead of the Normal (z)**
- CI for the population *mean* : $\overline{x} \pm \boxed{t_{n-1,1-\alpha/2}} \boxed{s} / \sqrt{n}$

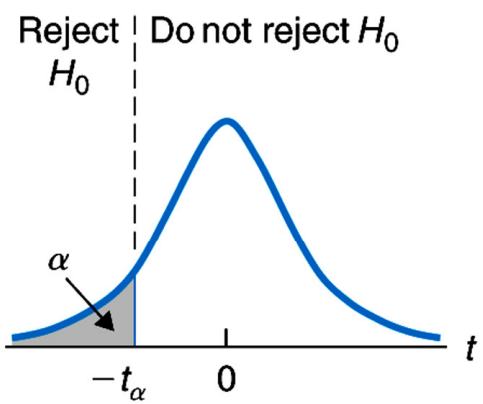
Hypothesis test : find the rejection region

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



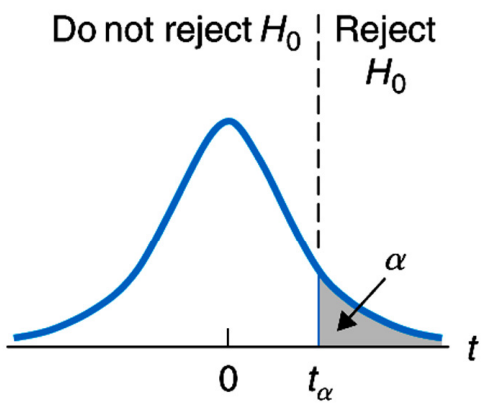
Two-tailed

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



Left-tailed

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



Right-tailed

Example

Example 9.1 Daily intake of energy (kJ) for 11 women :

5260 5470 5640 6180 6390 6515 6805 7515 7515 8230 8770

- Make a 95% for the mean daily energy intake (kJ) of the population of women ...
- Test the hypothesis that the mean is equal to the recommended value (7725 kJ) ...

Test

1

2

3

4

5

Test for comparing two (independent) means : equal variances

- We want to compare the means of two sets of measures :
  - Group 1 (p. ex. 'control') : $x_1, \ldots, x_n$
  - Group 2 (p. ex. 'treatment') : $y_1, \ldots, y_m$
- We can *model* these data as :

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

  where Δ signifies the effect of the treatment (compared to the 'control' group)
- $H : \Delta = 0$ vs. $A : \Delta \neq 0$ or $A : \Delta > 0$ or $A : \Delta < 0$

Equal variances, cont.

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

  $\hat{\Delta} = \bar{y} - \bar{x}$; $Var(\hat{\Delta}) = \dfrac{\sigma^2}{n} + \dfrac{\sigma^2}{m} = \dfrac{n+m}{nm}\sigma^2$
- We assume that :
  - the variances of the 2 samples are *equal* :
    $Var(\epsilon) = Var(\tau)$
  - the observations are *independent*
  - *the 2 samples are independent*
- We can estimate the variances *separately* :
  $s_x^2 = ((x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2)/(n-1)$
  $s_y^2 = ((y_1 - \bar{y})^2 + \cdots + (y_m - \bar{y})^2)/(m-1)$
- When the variances are *equal*, we can combine the two estimators : $s_p^2 = ((n-1)s_x^2 + (m-1)s_y^2)/(n+m-2)$

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

Test for comparing two (independent) means : unequal variances

- If $\sigma_x^2 \neq \sigma_y^2$, we can use

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

- The distribution of the statistic T_{Welch} *is only approximately* t, with a number of degrees of liberty calculated based on s_x, s_y, n and m
- Welch test
- In practice, if the variances are rather different (ratio more than 3), we could use this statistic (instead of the one with variance s_p^2)

Example

Example 9.2 Energy expenditure for groups of thin and obese women :

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				

- Test the hypothesis that the two population means are equal

...

Test

1

2

3

4

5

Paired experiments

- For an experiment carried out in *blocks of two units*, the *power* of the t-test can be increased
- This idea permits us to *eliminate the influences of other variables* (e.g. age, sex, *etc.*), in giving them different 'treatments'
- Thus, we have a *more precise* comparison of the two conditions

t-test for a paired experiment

- The data are of the form :

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

- *Each block* allows us to evaluate the effect of the treatment
- Here, we consider *the differences*

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

as a sample of measurements coming from a distribution with expected value Δ

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

Example 9.1, cont.

Example 2.2, cont. : Daily intake of energy of 11 women pre- and post-menopausal :

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

- Test the hypothesis that there is no difference in daily energy intake before and after menopause **...**

Test

1

2

3

4

5

Research process

- Scientific *question* of interest
- Decide *what data* to collect (and how)
- Collection and *analysis* of data
- Conclusions, generalizations : *inference* on the population
- *Communication* and dissemination of results

Generic question :
Does a 'treatment' have an 'effect' ?

Exemples :

- Does smoking cause cancer, heart disease, *etc* ?
- Does eating oat bran lower cholesterol ?
- Does échinacea prevent illness ?
- Does exercise slow the aging process ?

Types of studies

- A basic means to address this type of question involves *comparing two groups* of study subjects :
  - *Control group* : provides a baseline for comparison
  - *Treatment group* : group receiving the 'treatment'
- *Experimental study* : subjects assigned to groups by the investigator
  - *randomization* : protects against bias in assignment to groups
  - *'blind', 'double-blind'* : protects against bias in outcome assessment/measurement
  - *placebo* : artificial/fake treatment
- *Observational study* : subjects 'assign' themselves to groups
  - *confounder* : associated with both group membership/risk factor *and* with the outcome of interest

A few comments

- With a well-planned and well executed controlled experiment, it is possible to infer *causality*
- This is *not possible* with observational studies due to the presence of confounders
- With confounding, it is not possible to tell whether the observed difference between groups is due to the *treatment* or to the *confounding factor*
- Not always possible to carry out an experiment, for *pratical* and *ethical* reasons

Statistical models

- A **statistical model** is an approximate mathematical description of the mechanism that generated the observations, which takes into account *unexpected random errors* :
  - gives an *idealistic* representation of reality
  - makes *explicit assumptions* (that could be **false** ! !) about the process under study
  - permits an *abstract* reasoning
- The model is expressed by a Le modéle s'exprime par une *family of theoretical distributions* that contains the 'ideal' cases for the included RVs
  - e.g. : tosses of a coin **...**
- A useful model offers a *good compromise* between
  - *true* description of the reality (many parameters correct assumptions)
  - *ease* of mathematical manipulation
  - production of solutions/predictions *close* to the observation(s)

A simple model

A simple case : several measures of a physical quantity μ are taken, e.g. length of a field, person's height *...*

- Such measures possess in general a *random* component due to *measurement errors*
- One possible error mechanism :

$$\begin{array}{ccccc} \text{measure} & = & \text{true theoretical value} & + & \text{measurement error} \\ y & = & \mu & + & \epsilon \end{array}$$

- that is : measures with *additive errors*
- If there is no colitsystematic error (biais), the random error should be 'centered' ($E[\epsilon] = 0$)
- Often reasonable to think that *the precision* of each measure is *the same* ($Var(\epsilon) = \sigma^2$ for each measurement)
- *One possible specification* for the error distribution is *Normal* $N(0, \sigma^2)$
- **All models are wrong ; some are useful**

Estimation of the unknown parametres

- Once a model is chosen, we are interested in estimating unknowns : *the parameters of the model*
- We observe *realizations* of a RV for which the distribution is known (other than the parameter values)
- Thus, we must *estimate* the parameters using the observations $X_1, \ldots, X_n$
- $\hat{\mu} = \overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$
- $\hat{\sigma}^2 = S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$
- The estimator S^2 is *unbiased* for σ^2, and is *independent* of that for μ $(\overline{X})$

BREAK

Bivariate data

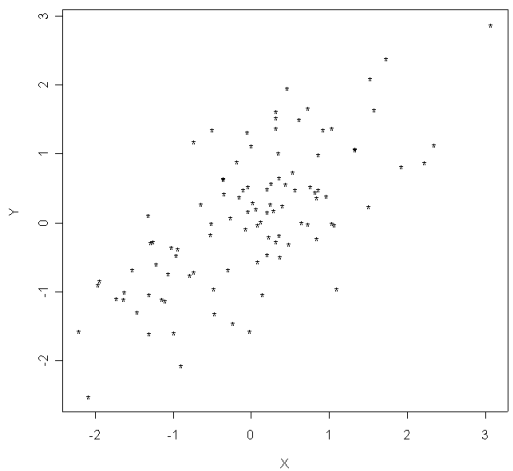
- Measures on *two* variables; e.g. X and Y
- We will consider the case of two *continuous* variables
- Want to discover the *relationship* between the two variables
  - forearm length and height
  - height and weight
  - expression of gene A and gene B
- We will consider datasets that are (at least approximately)

  bivariate normal $\Leftrightarrow$ oval-shaped

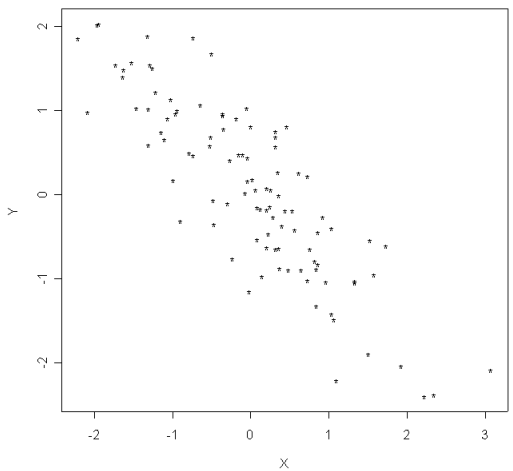
- $(X, Y) \sim BVN((\mu_x, \mu_y), (\sigma_x^2, \sigma_y^2), \rho)$

Exploratory analysis : scatterplot

- **Graphical summary** of a bivariate dataset using a *scatterplot* (or *cloud*)
- Values of one variable are plotted on the horizontal axis and values of the other on the vertical axis
- Can be used to see how values of 2 variables tend to move with each other (that is, how the variables are **associated**)

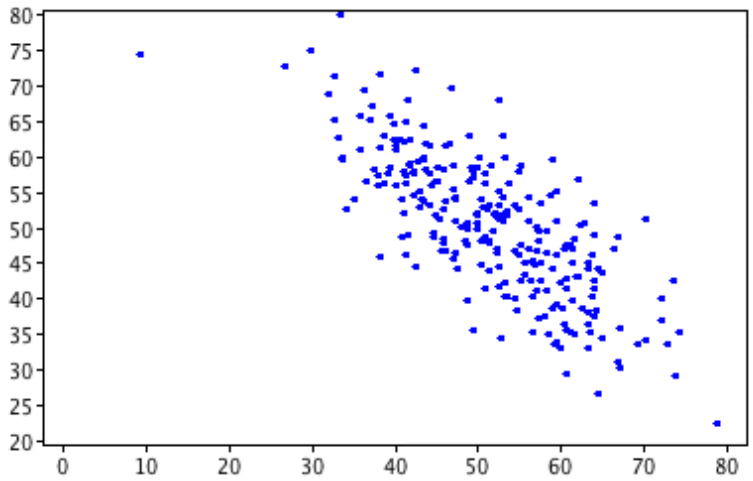


(a) positive association

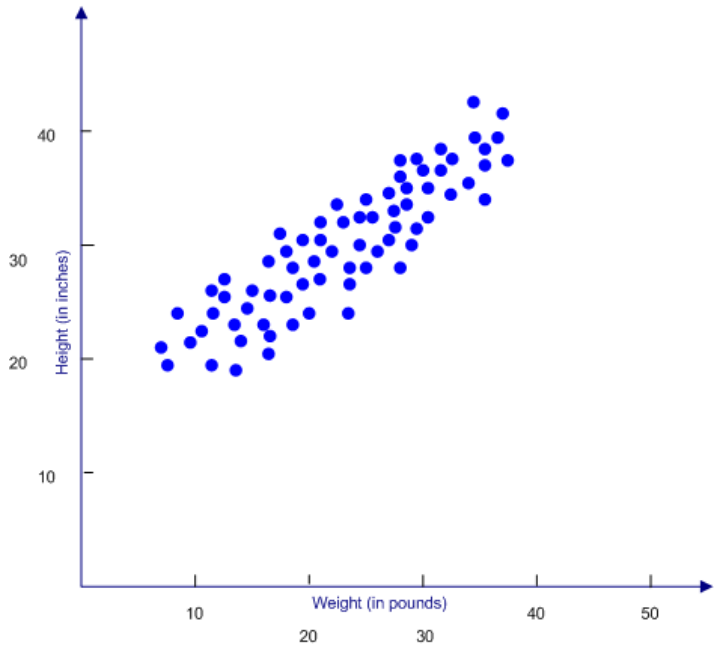


(b) negative association

Scatterplot



(a)



(b)

QCM : *What is the association between X and Y ??*

(a) none **(b)** positive **(c)** negative **(d)** impossible to determine

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

Numerical summaries

- Typically, bivariate data are summarized (numerically) with **5 statistics**
- These give a good summary for point clouds with the same general form that we just saw (oval)
- We can summarize each variable *separately* : $\overline{X}, s_X; \overline{Y}, x_Y$
- But these values don't say how the values of X and Y *vary together*

Correlation

- Let X and Y be RVs, with $Var(X) > 0, Var(Y) > 0$. The **correlation** $\rho(X, Y)$ is defined as :

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

- ρ is a *unitless quantity*, $-1 \leq \rho \leq 1$
- The correlation ρ, like the covariance, is **a measure of linear association** (the degree of linearity) of RVs X and Y
- Values of ρ close to 1 or -1 indicate strong linearity between X and Y, while values close to 0 indicate an absence of any **linear** relationship
- the sign of ρ indicates the direction of the association (positive or negative, corresponding to the slope of the line)
- When $\rho(X, Y) = 0$, X and Y are **uncorrelated**

Sample correlation coefficient

- The **sample correlation coefficient** r (or $\hat{\rho}$) is defined as the mean value of the (normalized) product XY :

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

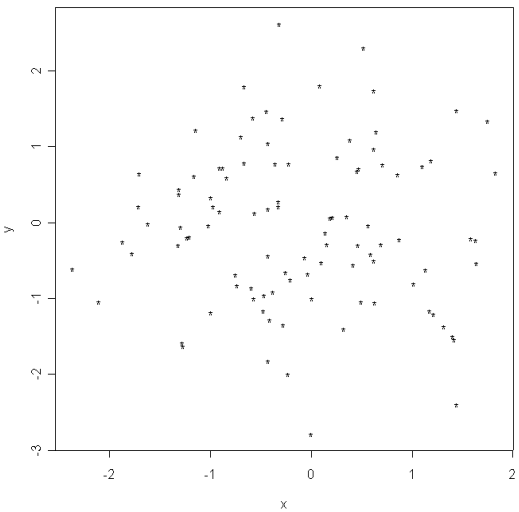
- centered-scaled = standardized (normalized)
  $= (X - \text{mean}(X))/\text{SD}(X)$
- r is a *unitless* quantity
- $-1 \leq r \leq 1$
- r is a measure of **LINEAR ASSOCIATION**

Correlation $\neq$ Causation

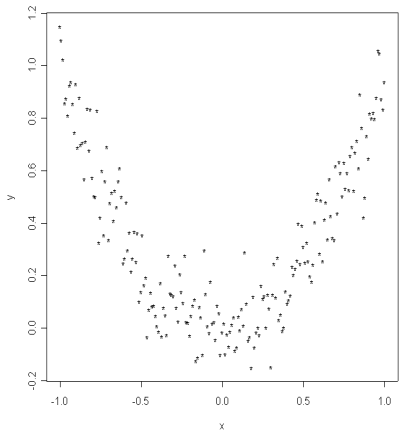
- We *cannot deduce* that, when X and Y are strongly correlated (r close to -1 or 1) that X *causes* a change in Y
- Y could be causing X
- X and Y could be varying with a third variable, perhaps an unknown factor (whether causal or not, often time)
  - polio and soft drinks
  - number of firefighters sent to a fire and amount of damage
  - Children who get tutored get worse grades than children who do not get tutored
- If $r \approx 0$, there is no **LINEAR ASSOCIATION**

  – this is **NOT** to say that there is *NO ASSOCIATION*
- We cannot deduce the form of the scatterplot based only on the value of r

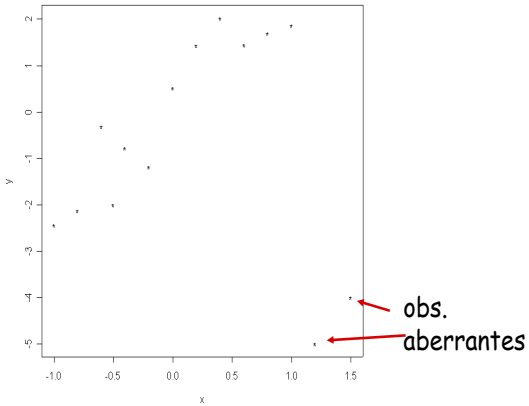
$r \approx 0$



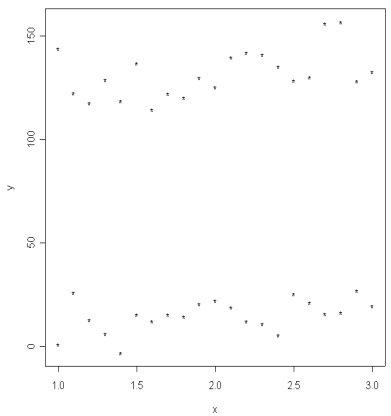
(a) random scatter



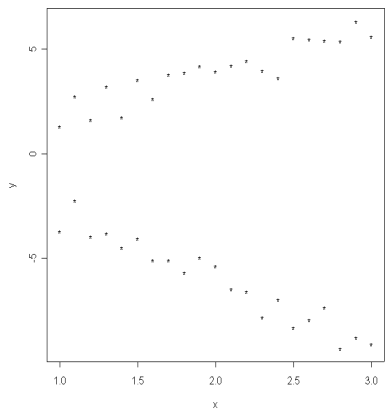
(b) curve



(c) outliers



(d) parallelism

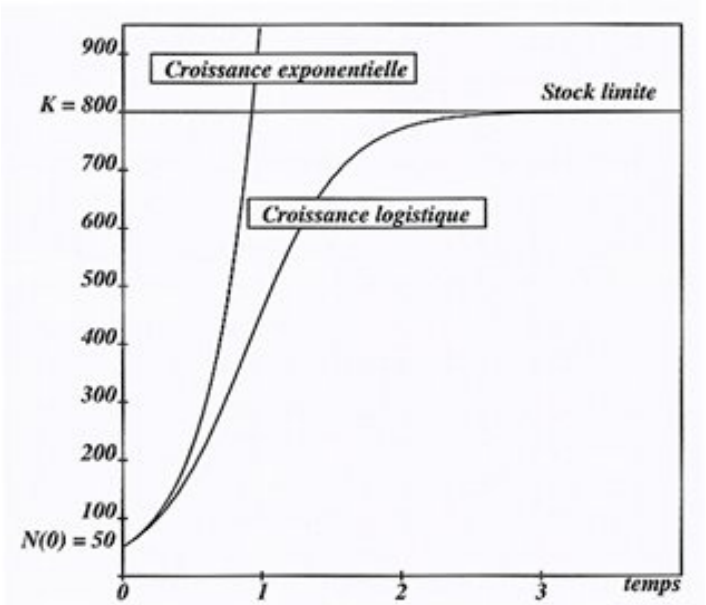


(e) two different lines

Modeling an oval-shaped point cloud

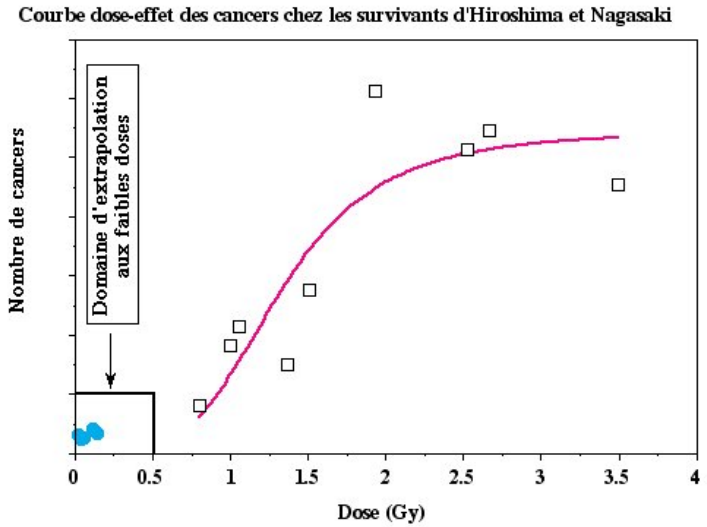
- Variable to be explained / response variable : Y
- Explanatory / predictor variables : X
  - The value of X is assumed to be known *without error*
  - We assume that variation in Y are *influenced* by X
  - The model expressed the assumed connection using a *mathematical relationship*
- Knowing these variables allows us to use the model to *predict* Y
  - Estimate the values of Y :
    – *pointwise*
    – using an *interval*
- The model also allows us to measure the *impact* (or *effect)* of an explanatory variable on Y

Deterministic or statistical relation



(a) deterministic

- *Only one value* of Y for a given value of X



(b) statistical

- *Multiple values* of Y for a given value of X
- *'Probabilize' Y* for a fixed value of X

Simple linear regression

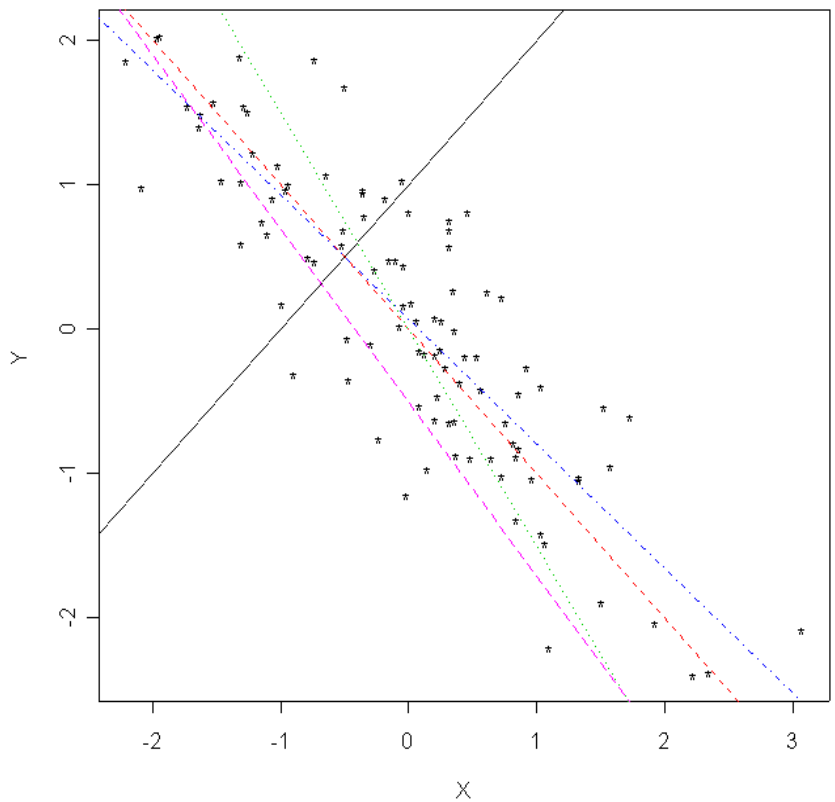
- Refers to drawing a (particular) line through a cloud of points
- Used for 2 objectives :
  - Explanation
  - Prediction
- Statistical linear model (linear in the **parameters**) :
  - $Y = \beta_0 + \beta_1 X + \epsilon \quad \Rightarrow \quad E[Y \mid X] = \beta_0 + \beta_1 X$
  - $E(\epsilon) = 0; \;\; Var(\epsilon) = \sigma^2$
- The equation to predict Y when a specific value x is known :

$$\widehat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

- $\beta_0 =$ the *intercept*; $\beta_1 =$ the *slope* (in the population)

Which line ?

- There are many lines that could be drawn through a cloud of points
- How to choose one ?



Prediction by regression

- We can make a prediction using *the regression line* :

  when X goes up by 1 (SD), the predicted value of Y goes up ** NOT by 1 (SD) **, but only by r (SD) (goes down if r is negative) :

  - $$\frac{\widehat{Y} - \overline{Y}}{s_Y} = \boldsymbol{r} \frac{X - \overline{X}}{s_X}$$

- This prediction can also be expressed as :

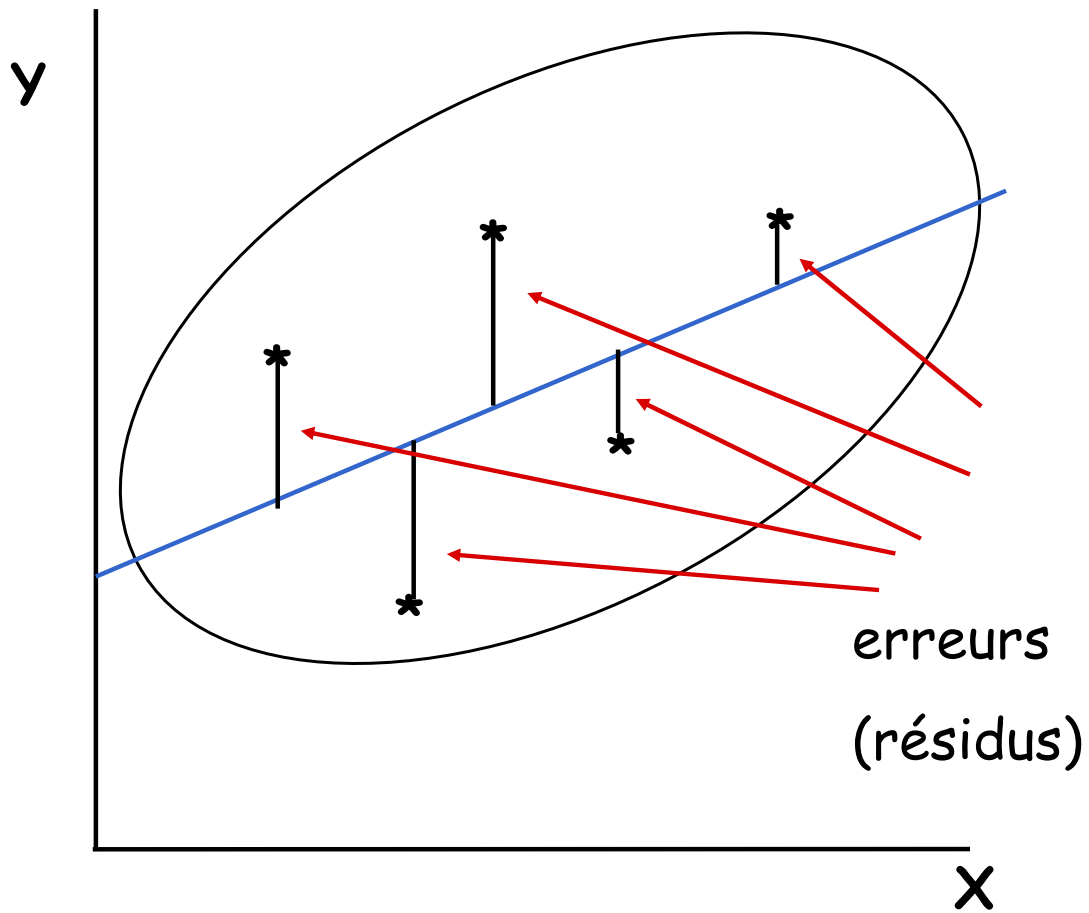
  $$\text{pred. } y = \text{int.} + \text{slope} \times x, \text{ with}$$

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

Least squares

Q : Where does this equation come from ?

A : It's the line that is 'best' in the sense that the sum of the *squared errors* in the vertical direction (Y) is a *minimum*



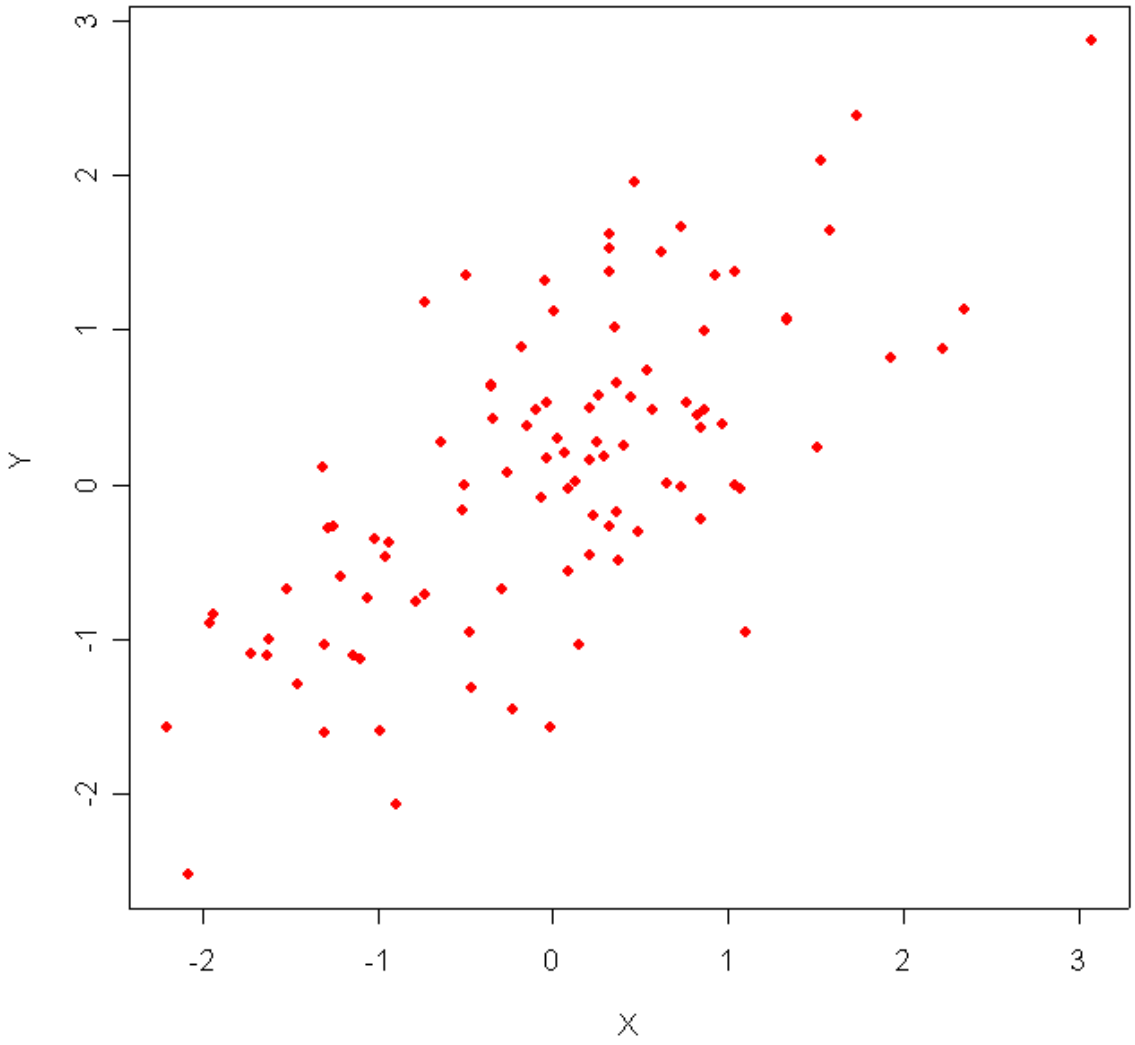
*** Interpretation of the parameters ***

- The equation of the regression line has 2 parameters : the *slope* and the *intercept*
- The *slope* is the mean (expected) change in Y for a change of 1 unit of X
- The *intercept* is the estimated value of Y when $X = 0$
- If the slope $= 0$, then X does not help in predicting Y (for a linear prediction)

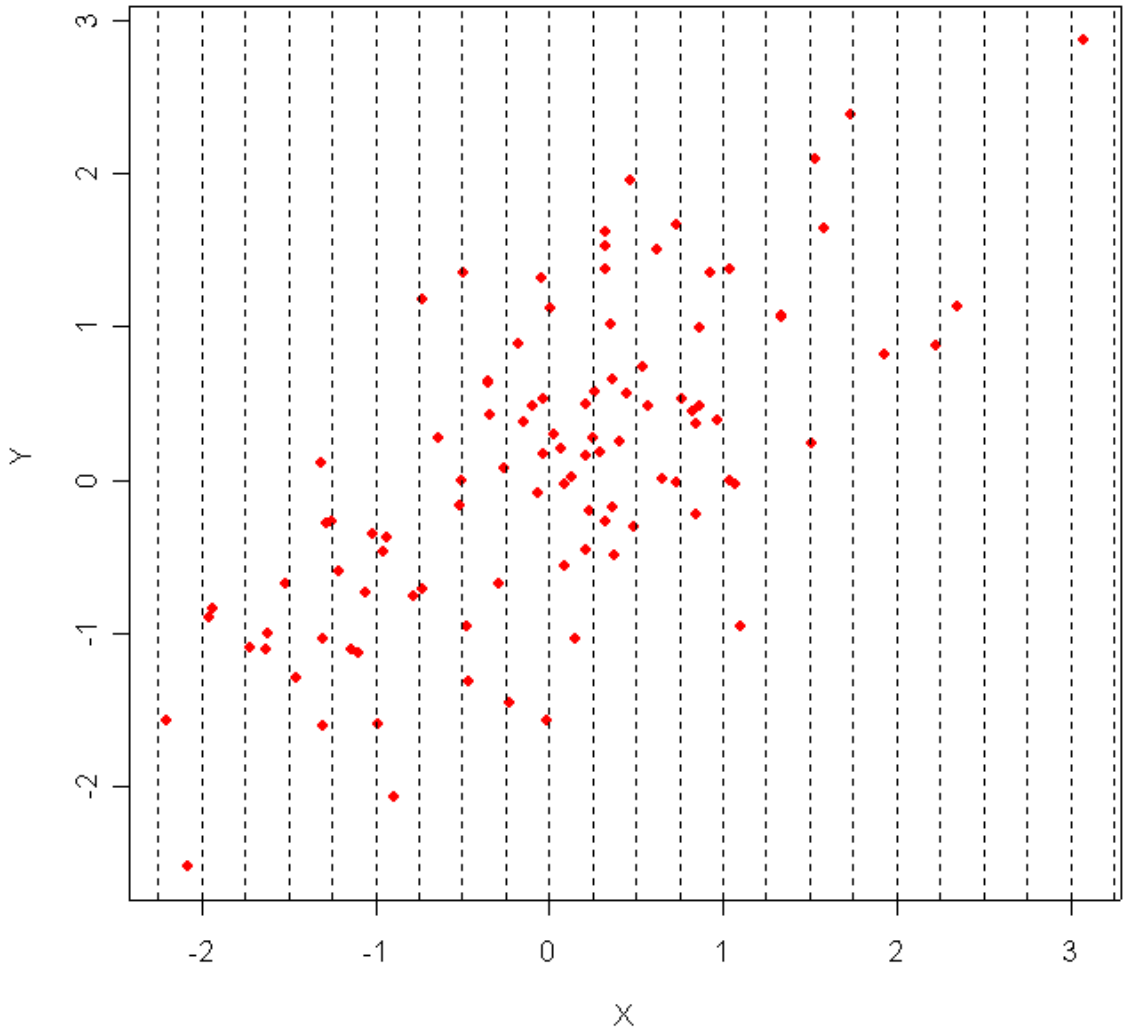
Another view of the regression line

- We can split the cloud of points into regions (*X-bands*) based on the values of X
- Within each X-band, mark the average value of Y (using only the values of Y whose X values are in that X-band)
- This is the *curve/graph of the means*
- The regression line can be considered as a *smoothed versione* of the curve of the means

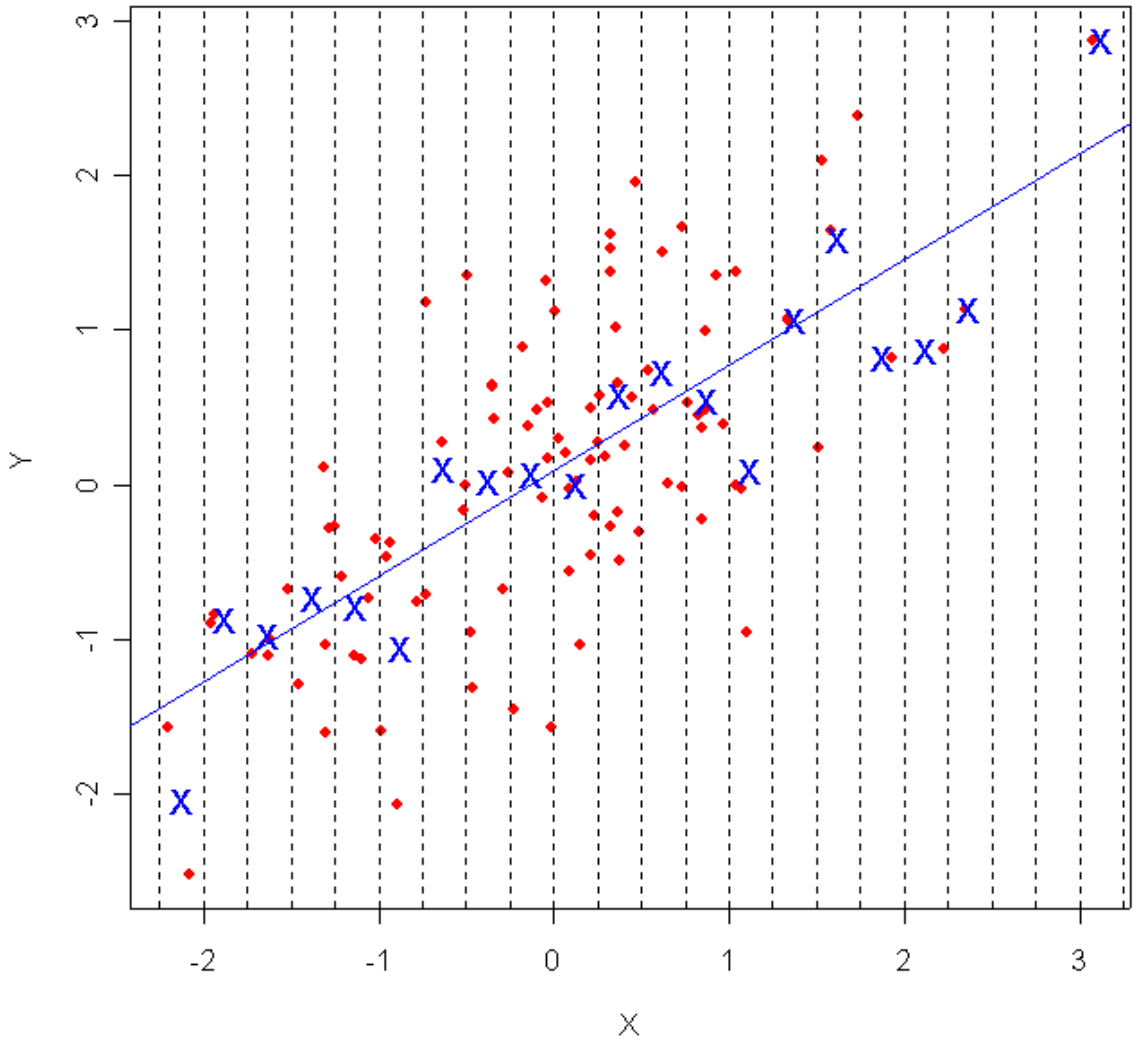
Scatterplot (again)



Creation of the X-bands



Graph of the means



Steps in regression

Starting with a sample of values for the response variable Y and the predictor variable(s) X :

- Verify the possibility of a linear relation between Y and X
  - graphical representation
  - correlation coefficient
- Parameter estimation
  - coefficients $\beta_i \Rightarrow \hat{\beta}_i$
  - standard deviation for the errors $\sigma \Rightarrow \hat{\sigma}$
- Model evaluation (next week)
  - measures of quality R^2, R^2_{adj}
  - global evaluation of model fit (Fisher's F)
  - test(s) for individual coefficients
  - examination of residuals, outlier detection, detection of influential points

Summary : simple linear regression (conceptual)

- For a scatterplot that is *oval-shaped*, we can find a line that serves to summarize the points
- A principle commonly used for fitting this line is *least squares* : the total of the squares of the (vertical) errors is minimized
- According to the principle, the regression prediction for Y knowing X tells us that :
  when X goes up by 1 SD_X, (the expected value of) Y goes up by rSD_X
- We can find the equation for the least squares line using the 5 statistics :

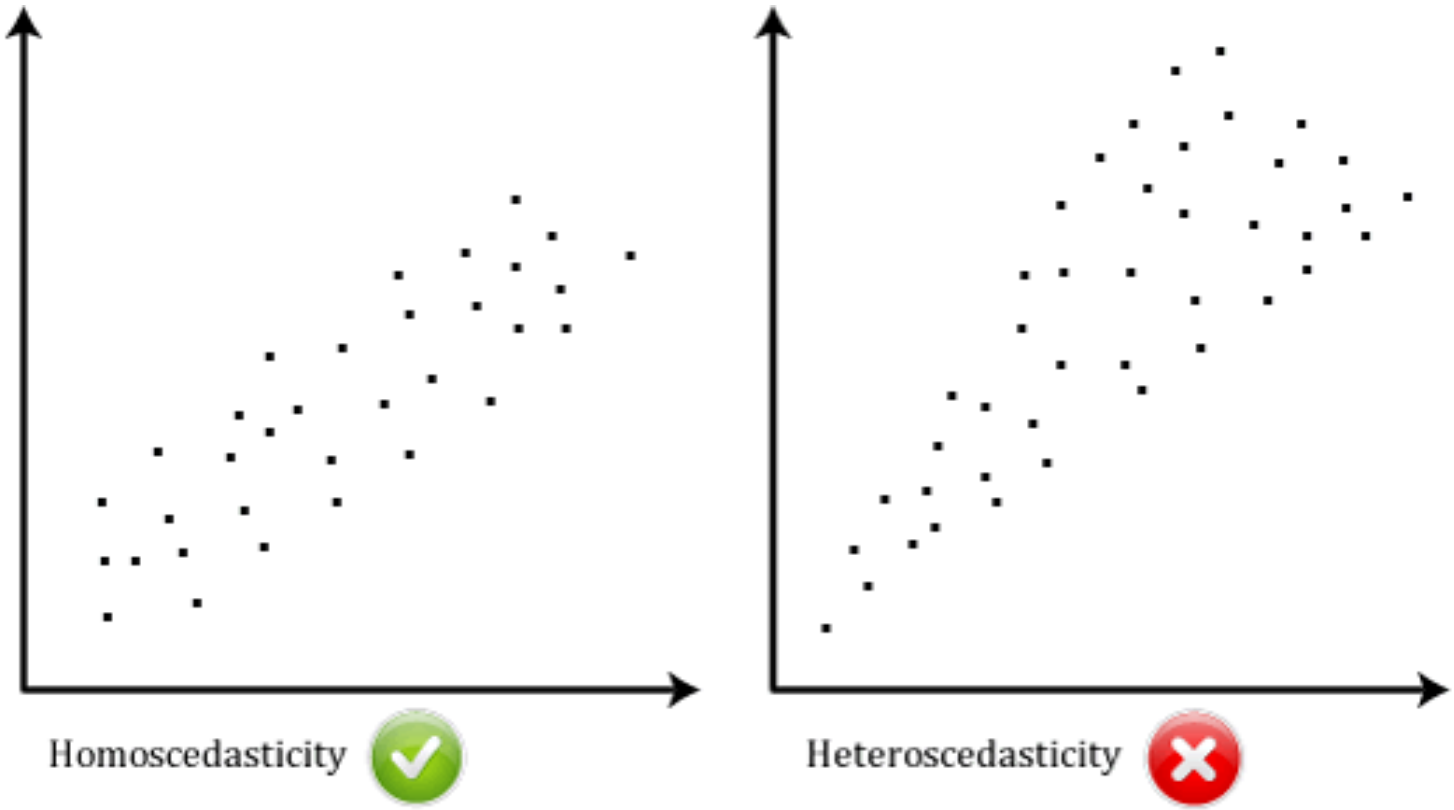
$$\overline{X}, SD(X), \overline{Y}, SD(Y), r$$

- The (estimated) *slope* is equal to $\hat{\beta}_1 = r\dfrac{s_Y}{s_X}$,
  the (estimated) *intercept* equals $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$

Simple linear regression – mathematical framework

- Here, we consider a model where the *response variable* y_i has a linear association with an *explanatory variable* (or *regressor* or *predictor*) x_i : $y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n$
- $\epsilon_1, \ldots, \epsilon_n$ are assumed to be random variables
  - uncorrelated
  - expectation $= 0$
  - variance $= \sigma^2$ for all $i = 1, \ldots, n$ (*homoscedastic*)
- x_i are assumed to be **constants** (measured without error)
- $\Rightarrow$ If the errors are also assumed to be *Normally distributed*, we can carry out *tests* and make *confidence intervals (CI)*
- Assumptions summarization :
  - **L**inear model (in the parameters)
  - **I**ndependent errors / observations
  - **N**ormal errors / observations
  - **E**qual error variances

Homoscedastic, heteroscedastic errors



Method of least squares (The details WILL NOT BE EXAMINED)

- The data are only a *sample* (and not the entire population)
- Thus, we must *estimate* the values of the parameters β_0 (intercept) et β_1 (slope) (as well as the error variance σ^2) :

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

- According to the *least squares principle*, we are looking for estimators that minimize :

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

- 'SS' = 'sum of squares'

Method of least squares, cont.

This is now *an optimization problem*, of finding the values $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize

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

To solve this, differentiate with respect to β_0, β_1 ; find the zeros :

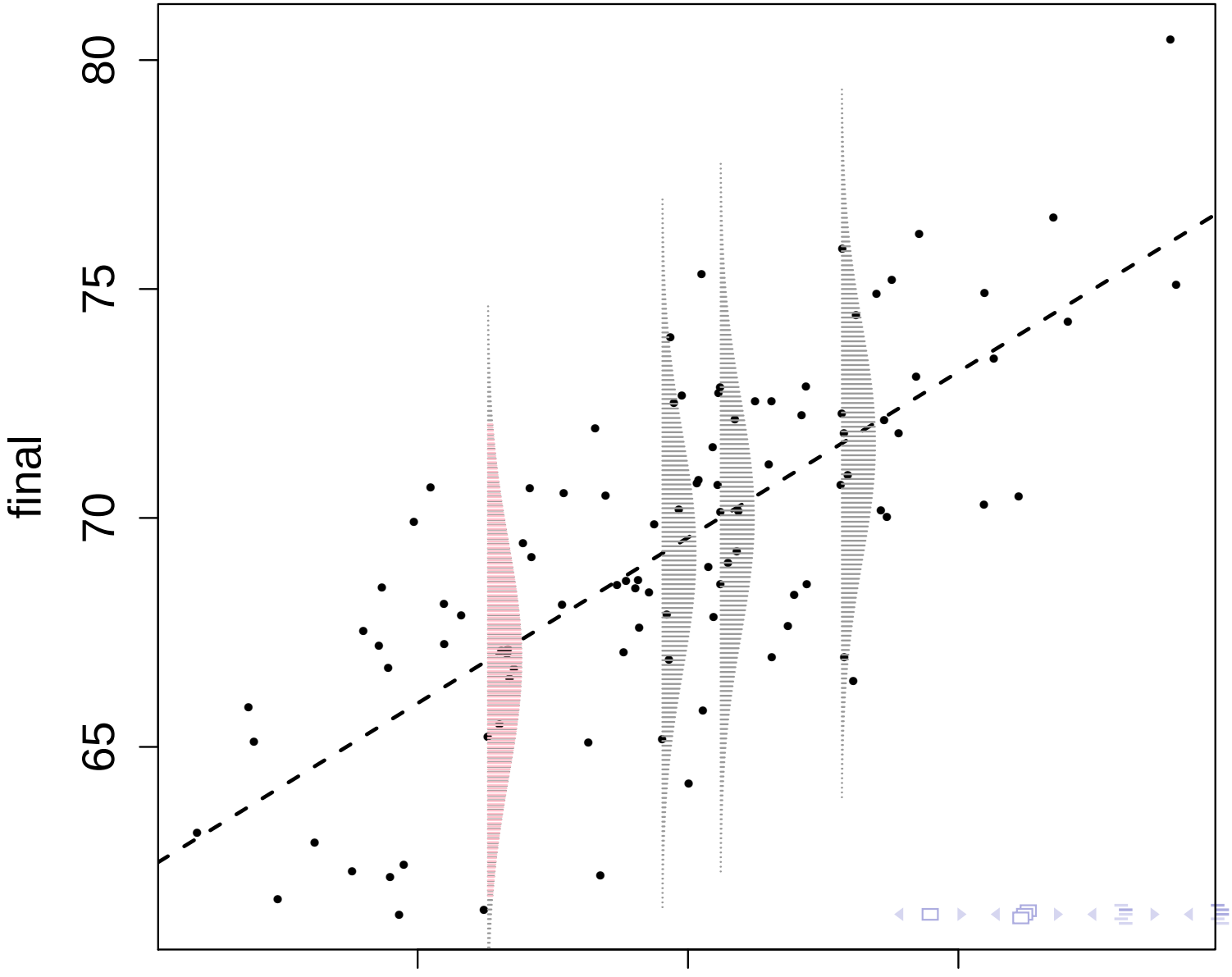
$$\begin{aligned}
\frac{d}{d\beta_0} &= \sum_{i=1}^{n} -2(y_i - \beta_0 - \beta_1 x_i) = 0 \\
&\Rightarrow \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \\
&\Rightarrow \sum_{i=1}^{n} y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_i = 0 \\
&\Rightarrow \sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n}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

Least squares, cont.

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

Simultaneous solution of (*) and (**) for the parameters β_0 et β_1 gives us the **regression estimates**.

Conditional Normal distribution : graphically



Conditional Normal distribution : algebraically

- The (univariate) *Normal distribution* depends on 2 parameters : the mean and the variance (equivalently, the SD)
- The *conditional expectation* is the regression prediction of Y given X : $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
- The *conditional error (RMSE)* is a new variability measure : the variability of the conditional expectation of Y knowing X, *i.e.*, the variability around the regression line ; it's the square root of the *mean square error (MSE)*
- MSE = arithmetic mean* of the squared deviations between the predictions and the observations
- *(instead of dividing by n, divide by *degrees of freedom*)
- $RMSE(Y) = s_Y \sqrt{(1 - r^2)}$

Properties of the estimator for the slope

- Estimated regression line : $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$
- The least squares estimate for the slope β_1 can be written as :

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

- Expected value of the estimator : $E[\hat{\beta}_1] = \beta_1$
- Variance of the estimator : :

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

- We need an estimator for σ^2 $(e_i = y_i - \hat{y}_i)$:

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

Test/confidence interval for the slope

- To test $H : \beta_1 = \beta_1^H$ against $A : \beta_1 \neq \beta_1^H$:

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

- We REJECT H if : $|t\text{-slope}_{obs}| > t_{\underline{n-2}, 1-\alpha/2}$
- The IC with level $1 - \alpha$ for the slope $]\beta_1$ est :

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