SPLEX
Statistiques pour la classification et fouille de données en génomique
Tests d'hypothèses

Pierre-Henri WUILLEMIN

Decision-axe IA-LIP6
pierre-henri.wuillemin@lip6.fr

Statistical hypothesis tests : introduction

Définition

An **hypothesis test** *is a decision rule used to determine which of two hypotheses concerning the value of a parameter (p, μ, σ^2, ...) is the more plausible.*

The first and perhaps most complicated step in constructing a hypothesis test is to identify the two hypotheses and formulating them in statistical language.

The two hypotheses to be confronted will always be noted :

- H_0 : null hypothesis and
- H_1 : alternative hypothesis

These two hypotheses must be mutually exclusive.

In principle, H_0 is the hypothesis we are trying to verify.

Statitstical tests

Formalisation

Let X follow a law P_θ on $\mathcal{X}$, parametrized by $\theta \in \Theta$. We have a sample $X_1, \cdots, X_n$, all i.i.d. of law P_θ.

Let $\theta = \theta_0 \cup \theta_1$ be a partition. On the sample, test the 2 hypotheses :

$$H_0 : \theta \in \theta_0 \qquad\qquad H_1 : \theta \in \theta_1$$

Exemple

In an assembly of 100 people, each person is asked to give a random number between 0 and 9. Note $x_i \in \{0, \cdots, 9\}$ the number given by individual i and n_j the number of individuals who gave the number j.
The results (i.e. the set of (j, n_j) where $j = 0, \cdots, 9$) are as follows :

$$(0, 10), (1, 8), (2, 9), (3, 14), (4, 8), (5, 9), (6, 11), (7, 9), (8, 12), (9, 10)$$

Can we consider that these numbers were indeed given at random, in the sense that the x_i are realizations of random variables i.i.d. distributed according to a uniform distribution on $\{0, \cdots, 9\}$?
The point is to test :

$$H_0 : X \text{ uniform over } \{0, \cdots, 9\} \qquad\qquad H_1 : \text{no}$$

Hypothesis testing in classic statistics

hypothesis

- Θ = set of values for the parameter θ
- Θ partitioned in Θ_0 et Θ_1
- *hypothesis* = a statement $H_0 =$ "$\theta \in \Theta_0$"and $H_1 =$ "$\theta \in \Theta_1$"
- H_0 = null hypothesis, H_1 = alternative hypothesis
- hypothesis H_i is *s*imple id Θ_i is a singleton ; otherwise it is *m*ultiple
- if $\Theta \subset \mathbb{R}$, unilateral test = values in Θ_1 are all bigger or smaller than values in Θ_0 ; otherwise it is bilateral.

	hypothesis	test
$H_0 : \mu = 4$ $H_1 : \mu = 6$	simple simple	unilateral
$H_0 : \mu = 4$ $H_1 : \mu > 4$	simple multiple	unilateral test
$H_0 : \mu = 4$ $H_1 : \mu \neq 4$	simple multiple	test bilatéral
$H_0 : \mu = 4$ $H_1 : \mu > 3$	simple composée	incorrect formulation : hypothesis not mutually exclusive.

Decision rule

- Constructing the decision rule means determining which values the parameter under study (e.g. $\overline{x}$) is unlikely to take in the sample if the hypothesis H_0 is true.
- It is necessary to analyze the distribution of the estimator of the parameter in the sample when H_0 is true and to determine a critical rcritical egion, or region of rejection of H_0, such that if the value taken by the estimator is in this region, it is unlikely that H_0 is true.

- The critical region must take into account the form of the counter-hypothesis so that the rejection of H_0 means that H_1 is a plausible choice.

- The decision rule is based on the sampling results.
- Sampling results are examined after the determination of the decision rule, not before.

The values of the parameter under the various assumptions must not be based on the observed result from the sample.

Critical regions for numerical hypothesis

Critical regions

hypothèses	*Décision rule*
$H_0 : \mu = \mu_0$ $H_1 : \mu > \mu_0$	«reject H_0 if $\overline{x} > c$», where c is bigger than μ_0
$H_0 : \mu = \mu_0$ $H_1 : \mu < \mu_0$	«rejec H_0 if $\overline{x} < c$», where c is smaller than μ_0
$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	«reject H_0 if $\overline{x} < c_1$ ou $c_2 < \overline{x}$», where c_1 and c_2 are respectively smaller and bigger than μ_0

Errors in decisions

Decision taken \ "Reality"	H_0 is "true"	H_1 is "true"
H_0 is not rejected	✓	bad decision. type II error β
H_0 is rejected	bad decision. type I error α	✓

α = type I risk
= probability of making a type I error
= probability of rejecting H_0 konwing that H_0 is true
= P(rejeter $H_0 | H_0$ is true),

β = type II risk
= probability of making a type II error
= probability de rejecting H_1 knowing that H_1 is true
= P(reject $H_1 | H_1$ is true).

Calculating α (1/2)

exemple

- sample of size 25
- Estimation of parameter μ for a variable $X \sim \mathcal{N}(\mu; 100)$
- hypothesis : $H_0 : \mu = 10 \quad H_1 : \mu > 10$

Assuming H_0 : $$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{\overline{X} - 10}{10/5} = \frac{\overline{X} - 10}{2} \sim \mathcal{N}(0; 1)$$

Assuming H_0 : unlikely that $\overline{X}$ deviates from μ by more than 2 standard deviations (σ) (p=4,56%)

$\Longrightarrow$ unlikely that $\overline{X} < 6$ ou $\overline{X} > 14$

$\Longrightarrow$ The critical region could then be «reject H_0 if $\overline{x} > 14$»

Calculating α (2/2)

Exemple

- sample of size 25
- Estimation of parameter μ for a variable $X \sim \mathcal{N}(\mu; 100)$
- hypothesis : $H_0 : \mu = 10 \quad H_1 : \mu > 10$
- critical region : «reject H_0 si $\overline{x} > 14$»

$$\begin{aligned}
\alpha &= P(\text{reject } H_0 | H_0 \text{ is true}) \\
&= P(\overline{X} > 14 | \mu = 10) \\
&= P\left(\frac{\overline{X} - 10}{2} > \frac{14 - 10}{2} \middle| \mu = 10\right) \\
&= P\left(\frac{\overline{X} - 10}{2} > 2\right) = 0,0228
\end{aligned}$$

in general, α is fixed and we look for the critical region

Power of a test

$$\alpha = P(\text{reject } H_0 | H_0 \text{ is true})$$

$$\beta = P(\text{reject } H_1 | H_1 \text{ is true})$$

α et β move in opposite directions to each other

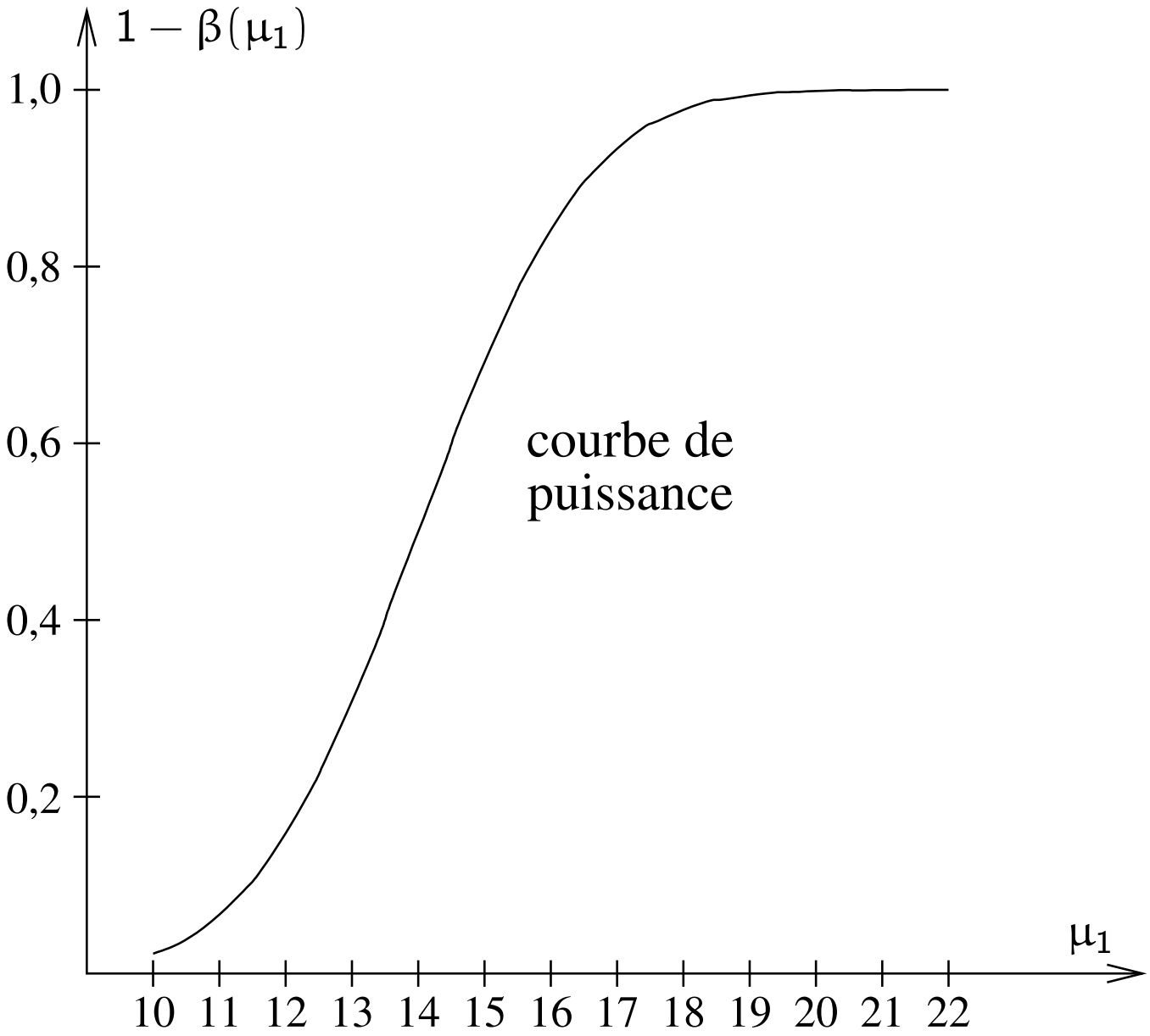
$\Longrightarrow$ test = trade-off between the two risks

H_0 = preferred hypothesis, verified to date and which we wouldn't would like to give up on.

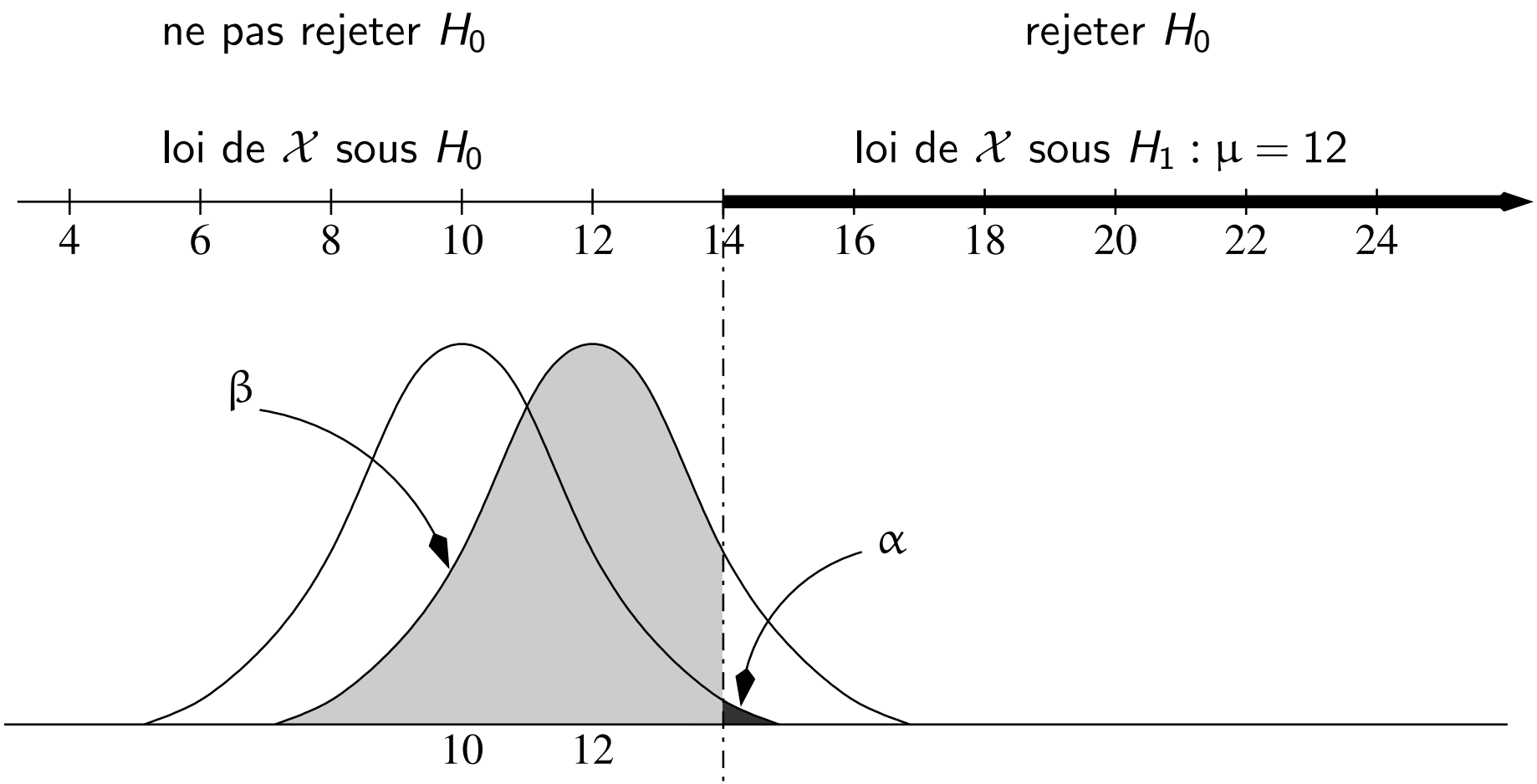
$\Longrightarrow$ α_0 is fixed :

- α must be $\leq \alpha_0$
- test minimising β under this constraint
- min β = max $1 - \beta$
- $1 - \beta$ = power of the test

Test Power Curve



Interpretation of α et β



Likelihood

We remember that :

$$P(X \mid Y) = \frac{P(Y \mid X) \cdot P(X)}{P(Y)}$$

Ou encore :

$$P(X \mid Y) \propto P(Y \mid X) \cdot P(X)$$

Let's denote θ the parameter we want to estimate and d the observation we make .

Définition (Likelihood, prior and posterior probabilities)

$$P(\theta \mid d) \propto P(d \mid \theta) \cdot P(\theta)$$

On nomme :

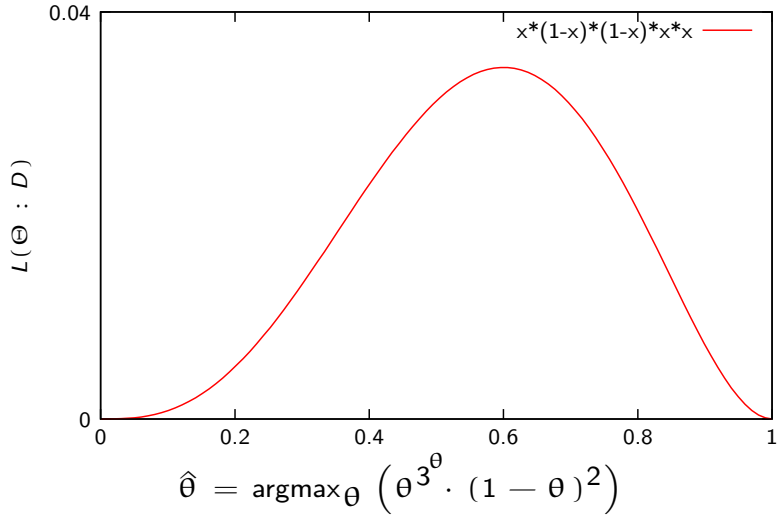
- $P(\theta)$ *prior distribution over* θ.
- $P(\theta \mid d)$ *posterior distribution over* θ.
- $P(d \mid \theta) = L(d, \theta) = L(\theta : d)$ *likelihood*.

Maximum Likelihood Estimation (MLE)

Let X be a binary variable. With $\theta = P(X = 1)$:

$$\Theta = \{\theta, 1-\theta\}$$
$$D = (1, 0, 0, 1, 1)$$
$$L(\Theta : D) = P(D \mid \Theta) = \prod_m P(X = d_m \mid \Theta)$$

Ici : $L(\Theta : D) = \theta \cdot (1-\theta) \cdot (1-\theta) \cdot \theta \cdot \theta$.



Probabilty estimator : frequency

For a sample where '1' appears p times, and '0' $q = n - p$ times :

$$L(\Theta : D) = \theta^p \cdot (1-\theta)^q$$

D'oé :

$$\frac{d(\Theta:D)}{d\theta} = p\theta^{p-1}(1-\theta)^q - q(1-\theta)^{q-1}\theta^p$$
$$\frac{d(\Theta:D)}{d\theta} = 0 \iff p(1-\theta) - q\theta = 0$$

finalement :

$$\hat{\theta} = \frac{p}{p+q}$$

Neyman-Pearson's Lemma

cas : $\Theta_0 = \{\theta_0\} \quad \Theta_1 = \{\theta_1\}$

Neyman-Pearson's Lemma

- there is always a most powerful (random) test for a given threshold given α_0
- it's a likelihood ratio

test :
$$\begin{cases} \frac{L(x,\theta_0)}{L(x,\theta_1)} > k \Rightarrow x \in A \text{ (accepter } H_0) \\ \frac{L(x,\theta_0)}{L(x,\theta_1)} < k \Rightarrow x \in W \text{ (rejeter } H_0) \\ \frac{L(x,\theta_0)}{L(x,\theta_1)} = k \Rightarrow \delta(x) = \rho \text{ (accepter } H_0 \text{ avec proba } 1-\rho \\ \qquad\qquad\qquad\qquad\qquad\qquad H_1 \text{ whit proba } \rho) \end{cases}$$

- k and ρ determined uniquely by $\alpha = \alpha_0$

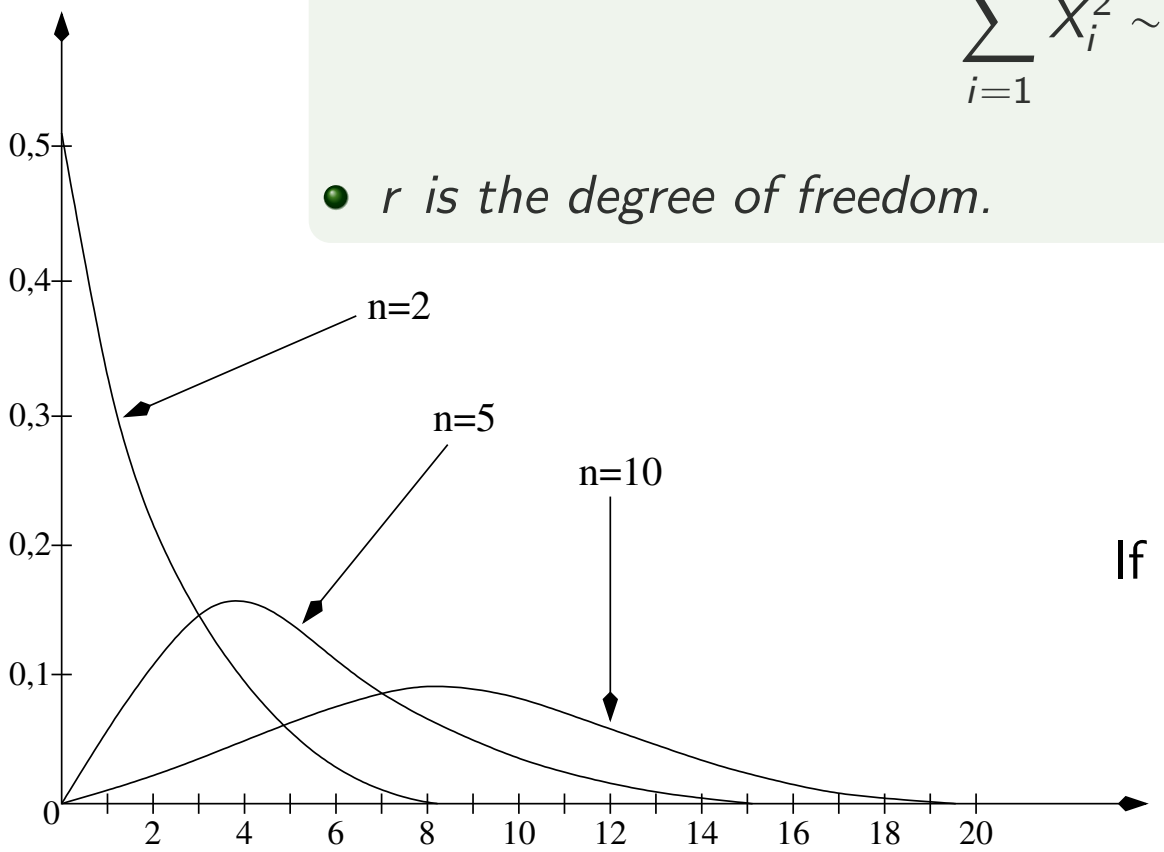
χ^2's distribution

If $(X_i)_{i\in\{1,\cdots,r\}}$ (i.i.d) $\sim \mathcal{N}(0;1)$ then $\sum_{i=1}^{r} X_i \sim \mathcal{N}(0;n)$

Définition (χ^2's distribution)

$$\sum_{i=1}^{r} X_i^2 \sim \chi^2_{(r)}$$

- *r is the degree of freedom.*
- *mean$=r$ and variance$=2r$*



If $r > 100$ then

$$\chi^2_{(r)} \approx \mathcal{N}(r;2r).$$

Confindence intervall for σ^2

If $(X_i)_{i\in\{1,\cdots,n\}}$ (i.i.d) $\sim \mathcal{N}(\mu, \sigma^2)$,
we know that $\overline{X} \sim \mathcal{N}(\mu; \sigma^2/n)$

Définition (Distribution of the corrected variance)

$$S^2 = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{n-1} \sim \chi^2_{(n-1)}$$

And then

Let X be a random variable with the distribution $\mathcal{N}(\mu; \sigma^2)$. Let s^2 the corrected variance observed on a sample of size n. Then the confidence interval for σ^2 with a confidence level $= 1 - \alpha$ isgiven by :

$$\left[\frac{(n-1)s^2}{c_{(n-1),\frac{\alpha}{2}}}; \frac{(n-1)s^2}{c_{(n-1),1-\frac{\alpha}{2}}}\right].$$

χ^2 : statistical test

Critical region for $H_0 : \sigma^2 = \sigma_0^2$

'H_1'	form of the critical region
$H_1 : \sigma^2 < \sigma_0^2$	$\frac{(n-1)s^2}{\sigma_0^2} < c_{n-1;1-\alpha}$
$H_1 : \sigma^2 > \sigma_0^2$	$\frac{(n-1)s^2}{\sigma_0^2} > c_{n-1;\alpha}$
$H_1 : \sigma^2 \neq \sigma_0^2$	$\frac{(n-1)s^2}{\sigma_0^2} < c_{n-1;1-\alpha/2}$ ou $\frac{(n-1)s^2}{\sigma_0^2} < c_{n-1;\alpha/2}$

Approximation of χ^2 by a gaussian

When $n > 30$:

$$c_{n;\alpha} \approx \frac{1}{2}\left[z_\alpha + \sqrt{2n-1}\right]^2$$

where z_α if the $1-\alpha$ quantile fo a random variable $Z \sim \mathcal{N}(0;1)$.

χ^2 : goodness-of-fit test

Définition

goodness-of-fit test

- *goodness-of-fit test = test whose outcome is the acceptance or rejection of the hypothesis that the observed sample follows a certain certain law.*
- *contre-hypothèse : ne précise pas de quelle autre loi il aurait pu être tiré.*

- population $\Longrightarrow$ répartie en k classes.
- hypothèse : répartition dans les classes connues.
  $\Longrightarrow p_l$ = proba qu'un individu appartienne à la classe l.
- on tire au hasard des individus dans la population.
- N_l = variable aléatoire «nombre d'individus tirés de classe l»
- soit $D^2_{(n)} = \sum_{l=1}^{k} \frac{(N_l - n.p_l)^2}{n.p_l}$.
  $D^2_{(n)}$ = écart entre théorie et observation
- $D^2_{(n)}$ tend en loi, lorsque $n \to \infty$, vers une loi du χ^2_{k-1}.

Utilisation de la loi du χ^2 : tests d'ajustement (2)

- population in k classes
- sample of size $n \Longrightarrow$ count $= (n_1, \ldots, n_k)$
- sample following the multinomial distribution $(p_1, \ldots, p_k)$
  $\Longrightarrow (n_1, \ldots, n_k) \approx (n.p_1, \ldots, n.p_k)$
- $D^2_{(n)} = \sum_{l=1}^{k} \frac{(N_l - n.p_l)^2}{n.p_l} \sim \chi^2_{k-1}$
- d^2 value of $D^2_{(n)}$ in the sample
  $\Longrightarrow$ if the sample follows $(p_1, \ldots, p_k)$ then d^2 is small
- In the table, d^2_α such that $P(\chi^2_{k-1} > d^2_\alpha) = \alpha$
- Decision rule : if $d^2 < d^2_\alpha$ then $(p_1, \ldots, p_k)$ is accepted as the distribution of the sample.

Independence test (1/3)

- 2 attributes X et Y
- classes for X : $A_1, A_2, \ldots, A_I$
- classes for Y : $B_1, B_2, \ldots, B_J$
- Sample of size n
- Contingency table :

$X \backslash Y$	B_1	B_2	$\cdots$	B_j	$\cdots$	B_J
A_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1J}
A_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2J}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
A_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{iJ}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
A_I	n_{I1}	n_{I2}	$\cdots$	n_{Ij}	$\cdots$	n_{IJ}

Independence test (2/3)

$X \backslash Y$	B_1	B_2	$\cdots$	B_j	$\cdots$	B_J	*total*
A_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1J}	$n_{1.}$
A_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2J}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
A_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{iJ}	$n_{i.}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
A_I	n_{I1}	n_{I2}	$\cdots$	n_{Ij}	$\cdots$	n_{IJ}	$n_{I.}$
total	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.j}$	$\cdots$	$n_{.J}$	n

$$\frac{n_{ij}}{n} = P(X \in A_i, Y \in B_j)$$

$$P(X \in A_i) = \frac{n_{i.}}{n} = \frac{\sum_{j=1}^{J} n_{ij}}{n} \qquad \text{and} \qquad P(Y \in B_j) = \frac{n_{.j}}{n} = \frac{\sum_{i=1}^{I} n_{ij}}{n}$$

$$X \text{ and } Y \text{ independants} \Longrightarrow P(X \in A_i, Y \in B_j) = P(X \in A_i) \times P(Y \in B_j)$$

Independence test (3/3)

$X \backslash Y$	B_1	B_2	$\cdots$	B_j	$\cdots$	B_J	*total*
A_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1J}	$n_{1.}$
A_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2J}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
A_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{iJ}	$n_{i.}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
A_I	n_{I1}	n_{I2}	$\cdots$	n_{Ij}	$\cdots$	n_{IJ}	$n_{I.}$
total	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.j}$	$\cdots$	$n_{.J}$	n

X and Y independants $\Longrightarrow \frac{n_{ij}}{n} = \frac{n_{i.}}{n} \times \frac{n_{.j}}{n} \Longrightarrow n_{ij} = \frac{n_{i.} \times n_{.j}}{n}$

$$\sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(n_{ij} - \frac{n_{i.} n_{.j}}{n}\right)^2}{\frac{n_{i.} n_{.j}}{n}} \sim \chi^2_{(I-1)\times(J-1)}$$

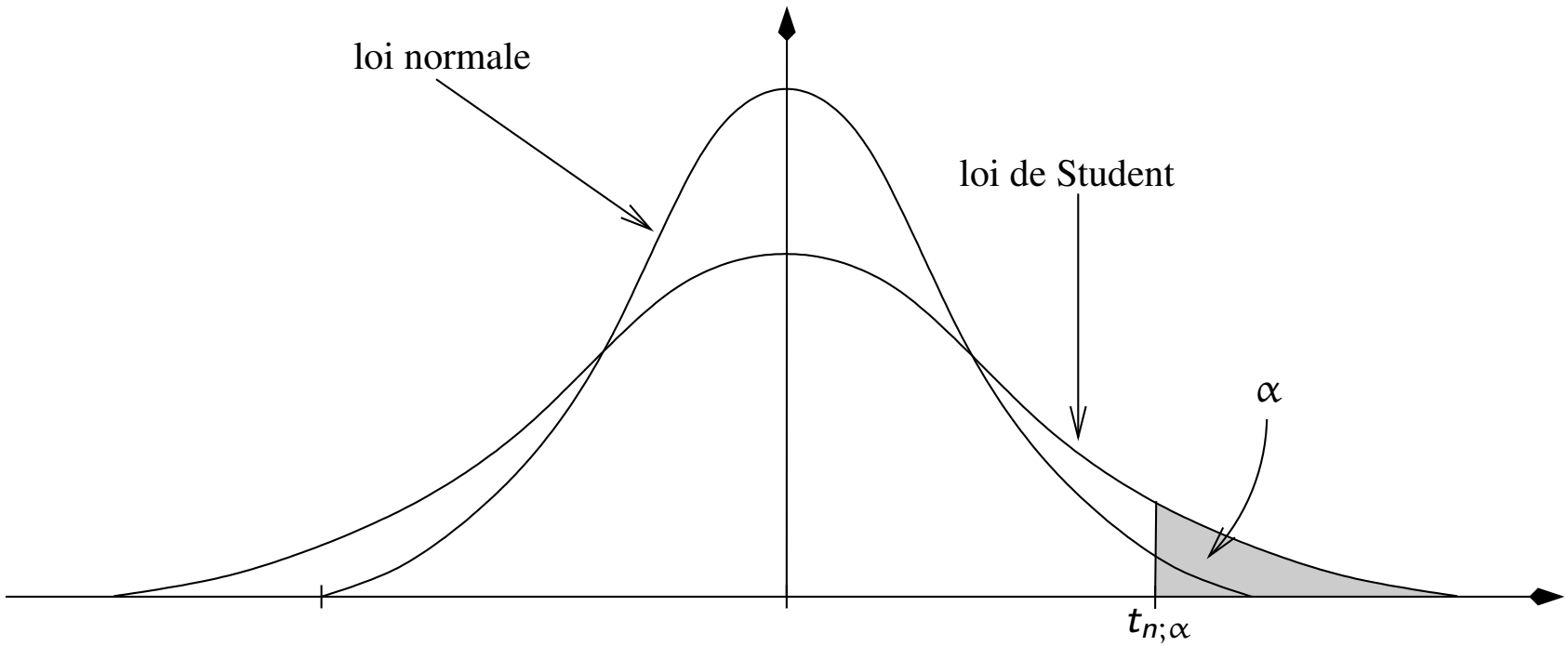
Student's distribution

Student's distribution

Student's distribution has only one parameter n (degree of freedom). The distribution with n degrees is T_n.

- If X follows T_n then $E(X) = 0$,
- $V(X) = \frac{n}{n-2}$ for $n > 2$.

. WHen $n >>$, T_n is close to $\mathcal{N}(0;1)$.



confidence intervall for μ

Let $(x_i)_{i \leq n}$ be a sample and $\overline{x}$ its mean and s^2 its variance.

Confidence intervall for μ with a confidence level $1 - \alpha$

Situation	Distribution	intervall
σ^2 known $X \sim \mathcal{N}$ where n big (> 75)	$\frac{\mathcal{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0;1)$	$\overline{x} \pm z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$
σ^2 unknown n bigger (> 75)	$\frac{\mathcal{X} - \mu}{S/\sqrt{n}} \sim \mathcal{N}(0;1)$	$\overline{x} \pm z_{\alpha/2} \times \frac{s}{\sqrt{n}}$
σ^2 unknown $X \sim \mathcal{N}$	$\frac{\mathcal{X} - \mu}{S/\sqrt{n}} \sim T_{n-1}$	$\overline{x} \pm t_{n-1;\alpha/2} \times \frac{s}{\sqrt{n}}$

Comparison of different samples

Assume 2 samples :

- n_1, $\overline{X}_1$, s_1^2 from a population μ_1, σ_1^2,
- n_2, $\overline{X}_2$, s_2^2 from another population μ_2, σ_2^2,

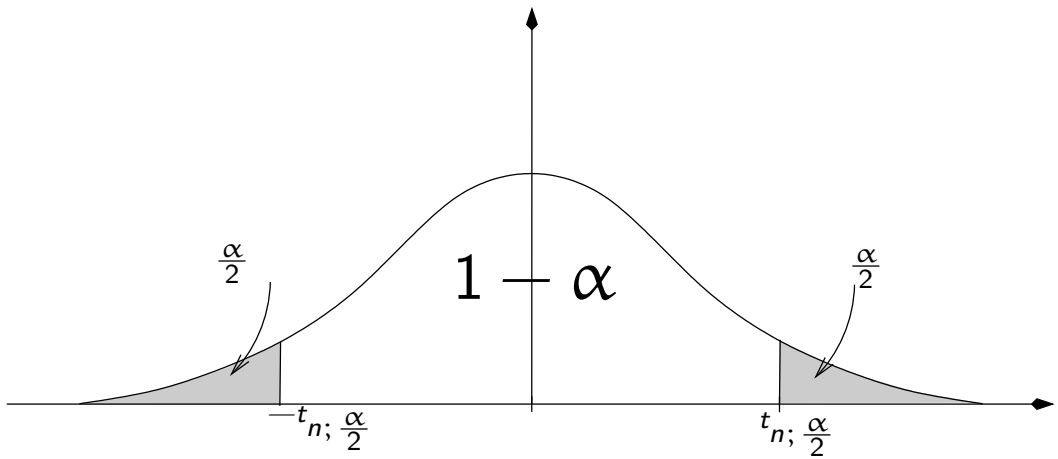
Comparison of μ_1 and μ_2

$H_O : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

Decision rule

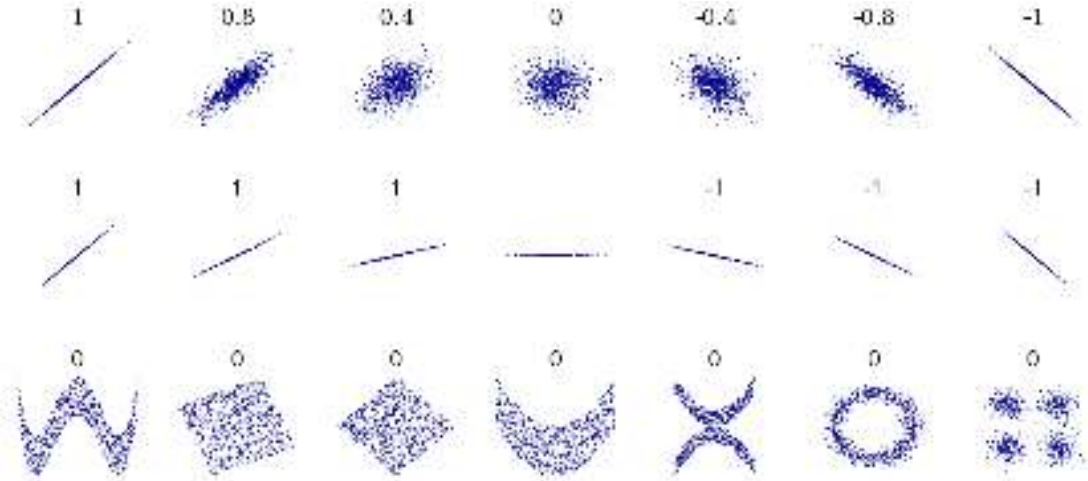
$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim T_{n_1+n_2-2} \qquad \text{with } s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n2 - 2}$$



Other tests in the TME

- (Pearson's) correlation coefficient

$$\rho_{XY} = \frac{cov(X,Y)}{\sigma_X \cdot \sigma_Y}$$



- Kruskal-Wallis (non parametric test)

Different salaries ?

Women : 23K, 41K, 54K, 66K, 78K.
Men : 45K, 55K, 60K, 70K, 72K.
Minorities : 18K, 30K, 34K, 40K, 44K.

$$H = \frac{12}{n \cdot (n+1)} \sum_{i=1}^{C} \frac{R_j^2}{n_j} - 3(n+1)$$

*H*0 : **no différence between différentes classes**
H=6.72
df=3-1=2
$\Rightarrow \chi^2_{2,0.05} = 9.48 > 6.72$
$\Rightarrow$ we cannot reject *H*0