Data analysis and stochastic modeling

Lecture 9 – Hypothesis testing

Guillaume Gravier

guillaume.gravier@irisa.fr

What for?

Hypothesis testing = make some decision on whether something is true or not based on experimental evidences, yet knowing the risk we are taking with that decision.

Typical hypotheses we want to test are:

- the mean time to failure of a system exceeds a threshold θ_0 (or not)
- the job arrival rate in a queing system is equal to λ_0 (or not)
- the distribution of some samples is Gaussian (or not)
- two estimated mean values correspond to the same mean (or not)
- an observed arrival process is Poisonian (or not)
- etc.

A useful reminder

$$\begin{aligned} &\text{Empirical mean estimator} \quad && \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \\ &\text{Empirical variance estimator} \quad && S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \end{aligned} \tag{1}$$

The central limit theorem states that

$$\frac{\overline{X} - m}{\sigma/\sqrt{n}} \quad \xrightarrow{\mathcal{L}} \quad \mathcal{N}(0,1)$$

Note on Gaussian variables: for Gaussian variables, it's not a convergence, it's an equality!

The rainmakers example

- rain levels (in mm) are assumed to be Gaussian $\mathcal{N}(600, 100)$
- people claim they can increase rain levels by 50 mm and, using their method over 9 years, we observed the following rain levels

year	1	2	3	4	5	6	7	8	9
mm	510	614	780	512	501	534	603	788	650

- Is this a scam or not?

Two hypotheses are confronted

$$\begin{array}{ll} H_0 & m = 600\text{mm} \\ H_1 & m = 650\text{mm} \end{array}$$

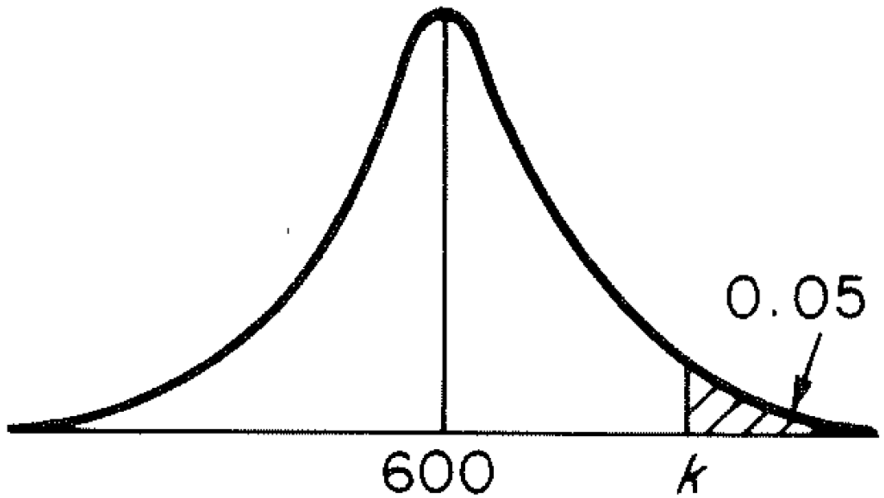
Since the method to increase rain levels is expensive, we want to use it only if we are pretty sure that it works, i.e. if we have only a 5 % chance of wrongly accepting H_1 from the evidences (risk $\alpha = 0.05$).

The rainmakers example (cont'd)

We study the empirical mean $\overline{X}$ over the nine year period

- if H_0 is true, $\overline{X} \rightsquigarrow \mathcal{N}(600, 100/\sqrt{9})$ [see lecture 4]
- decision rule
  - if $\overline{X} > k$, accept H_1 (or reject H_0)
  - if $\overline{X} < k$, reject H_1 (or accept H_0)
  - k is determined so that the probability of wrongly accepting H_1 is 0.05

$$k = 600 + \frac{100}{\sqrt{9}} 1.64 = 655$$



Conclusion: those rainmakers are ripping you off!

The rainmakers example (cont'd)

What if rainmakers were very unlucky over those nine years?

- if they were right, $\overline{X} \rightsquigarrow \mathcal{N}(650, 100/\sqrt{9})$
- an error is made each time $\overline{X} < 655$
- the probability of wrongly accepting H_0 (or wrongly rejecting H_1) is given by the risk β

$$\beta = P\left[U < \frac{655 - 650}{100/\sqrt{9}}\right] = 0.56$$

- H_1 defines the shape of the critical decision region $(650 > 600)$ but the threshold k only depends on H_0 and α
- additional knowledge on H_1 is used to compute the risk β

Concepts and vocabulary

- a **statistical hypothesis** is an assertion which can be valid or not
- a **statistical test** is a procedure to make a decision
- the **null hypothesis** H_0 is a claim that we are interested in accepting or rejecting
- the **alternative hypothesis** H_1 is the contradiction of H_0
- the **critical or rejection region** is the region where we reject H_0 (above k in the example) as opposed to the **acceptance region**
- wrongly rejecting H_0 is known as the **type 1 error** and the probability α that this happens is the **level of significance** of the test
- wrongly rejecting H_1 is a **type 2 error** and the quantity $1 - \beta$ is the **power** of the test

Hypotheses and types of errors

- α = probability of choosing H_1 when H_0 is true
- β = probability of choosing H_0 when H_1 is true

truth \ decision	H_0	H_1
H_0	$1-\alpha$	α
H_1	β	$1-\beta$

In practice, α is given by the decision maker and H_0 corresponds to the following

- a well established hypothesis, not contradicted so far
- a safe decision

    e.g. when testing a vaccine, H_0 is the less favorable hypothesis

- the only hypothesis that is easy to formulate

    e.g. $m = m_0$ is easir to formulate than $m \neq m_0$

UNIVERSITÉ DE RENNES 1

Methodology

1. define H_0 and H_1
2. determine the variables on which to make the decision
3. determine the shape of the critical region based on H_1
4. compute the critical region given α
5. eventually compute the power of the test

6. compute the experimental value of the decision variable
7. accept or reject H_0

Methodology

Many flavors and variants of tests:

- comparing data with theoretical distribution (fitting)

  $\rightarrow$ mean value, χ^2, etc.

- comparing two populations

  $\rightarrow$ χ^2, paired t-test, ranks, etc.

- likelihood ratio tests

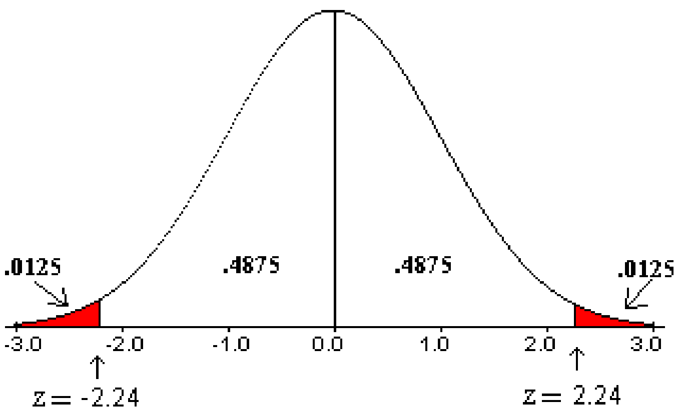
UNIVERSITÉ DE RENNES 1

Compare sample to mean value

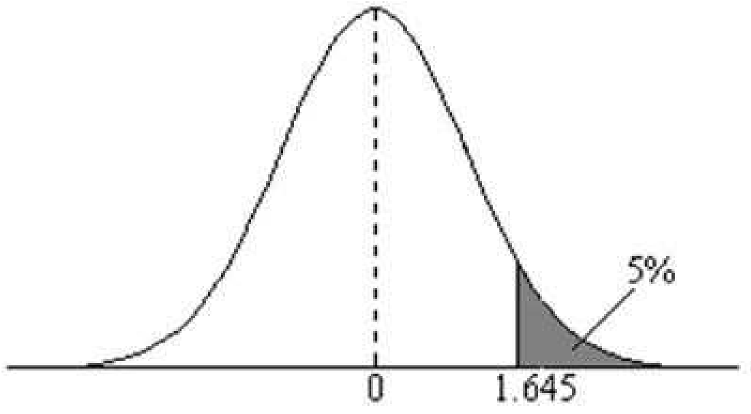
Given a set of evidence data $X_i = x_i$ i.i.d of unknown mean μ and (known) standard deviation σ, we might we to test the following based on the test statistics

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \rightsquigarrow \mathcal{N}(0,1)$$

$\mu = \mu_0$ vs. $\mu \neq \mu_0$
reject if $|Z| > k_\alpha$



$\mu = \mu_0$ vs. $\mu > \mu_0$
reject if $Z > k_\alpha$

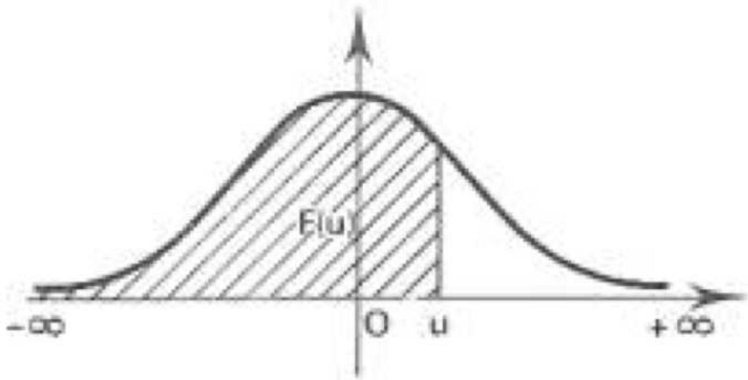


(a) One-tailed test

UNIVERSITÉ DE RENNES 1

Probability tables to determine the threshold

$$u_i = \frac{x_i - \bar{x}}{\sigma}.$$



L'emploi de cette table exige par conséquent la standardisation préalable de la valeur de X dont on veut connaître la probabilité cumulée; u se lit dans la première colonne pour sa partie entière et sa première décimale, la deuxième décimale se trouvant dans la première ligne.

u	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,500 0	0,504 0	0,508 0	0,512 0	0,516 0	0,519 9	0,523 9	0,527 9	0,531 9	0,535 9
0,1	0,539 8	0,543 8	0,547 8	0,551 7	0,555 7	0,559 6	0,563 6	0,567 5	0,571 4	0,575 3
0,2	0,579 3	0,583 2	0,587 1	0,591 0	0,594 8	0,598 7	0,602 6	0,606 4	0,610 3	0,614 1
0,3	0,617 9	0,621 7	0,625 5	0,629 3	0,633 1	0,636 8	0,640 6	0,644 3	0,648 0	0,651 7
0,4	0,655 4	0,659 1	0,662 8	0,666 4	0,670 0	0,673 6	0,677 2	0,680 8	0,684 4	0,687 9
0,5	0,691 5	0,695 0	0,698 5	0,701 9	0,705 4	0,708 8	0,712 3	0,715 7	0,719 0	0,722 4
0,6	0,725 7	0,729 0	0,732 4	0,735 7	0,738 9	0,742 2	0,745 4	0,748 6	0,751 7	0,754 9
0,7	0,758 0	0,761 1	0,764 2	0,767 3	0,770 4	0,773 4	0,776 4	0,779 4	0,782 3	0,785 2
0,8	0,788 1	0,791 0	0,793 9	0,796 7	0,799 5	0,802 3	0,805 1	0,807 8	0,810 6	0,813 3
0,9	0,815 9	0,818 6	0,821 2	0,823 8	0,826 4	0,828 9	0,831 5	0,834 0	0,836 5	0,838 9

UNIVERSITÉ DE RENNES 1

Compare sample to mean value (cont'd)

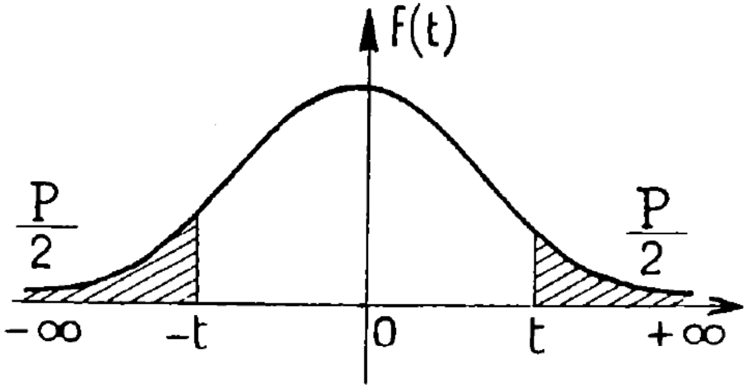
If standard deviation σ is unknown, Z can no longer be used and we use the Student test statistics

$$T = \sqrt{n-1}\frac{\overline{X} - m}{S} \rightsquigarrow t_{n-1}$$

Example:

- $H_0 \quad m = 30$ vs. $H_1 \quad m \neq 30$
- 15 samples with $\overline{x} = 37.2$ and $s = 6.2$
- under H_0, $t = \sqrt{14}\dfrac{37.2 - 30}{6.2} = 4.35$
- critical value at $\alpha = 0.05$ for T_{14} = 1.761 $\Rightarrow$ REJECT H_0!

Probability tables



ν \ P	0,90	0,80	0,70	0,60	0,50	0,40	0,30	0,20	0,10	0,05	0,02	0,01	0,001
1	0,158	0,325	0,510	0,727	1,000	1,376	1,963	3,078	6,314	12,706	31,821	63,657	636,619
2	0,142	0,289	0,445	0,617	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925	31,598
3	0,137	0,277	0,424	0,584	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841	12,929
4	0,134	0,271	0,414	0,569	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604	8,610
5	0,132	0,267	0,408	0,559	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032	6,869
6	0,131	0,265	0,404	0,553	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707	5,959
7	0,130	0,263	0,402	0,549	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499	5,408
8	0,130	0,262	0,399	0,546	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355	5,041
9	0,129	0,261	0,398	0,543	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250	4,781
10	0,129	0,260	0,397	0,542	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169	4,587

UNIVERSITÉ DE RENNES 1

The χ^2 fitting tests

X is a random variable divided into k classes of resp. probabilities $p_1, p_2, \ldots, p_k$ and we observe a sample of the variable with population size in each class $N_1 = n_1, N_2 = n_2, \ldots N_k = n_k$

Note that $E[N_i] = np_i$

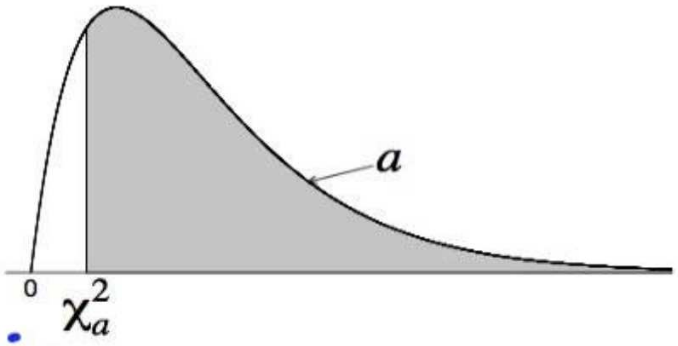
We want to test if the underlying law of the observed process fits the theoretical distribution defined by the p_i's (H_0) or not.

- test statistics is $D^2 = \displaystyle\sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i}$
- asymptotic distribution of the test statistics = χ^2_{k-1}

Notes:

- if m parameters are estimated from the sample (e.g. λ in Poisson) $D^2 \rightsquigarrow \chi^2_{k-1-m}$
- need for at least 5 (or 3) elements per class

χ^2 table



df	$\chi^2_{0.9995}$	$\chi^2_{0.999}$	$\chi^2_{0.995}$	$\chi^2_{0.990}$	$\chi^2_{0.975}$	$\chi^2_{0.95}$	$\chi^2_{0.90}$	$\chi^2_{0.85}$	$\chi^2_{0.80}$
1	0.000	0.000	0.000	0.000	0.001	0.004	0.016	0.036	0.064
2	0.001	0.002	0.010	0.020	0.051	0.103	0.211	0.325	0.446
3	0.015	0.024	0.072	0.115	0.216	0.352	0.584	0.798	1.005
4	0.064	0.091	0.207	0.297	0.484	0.711	1.064	1.366	1.649
5	0.158	0.210	0.412	0.554	0.831	1.145	1.610	1.994	2.343
6	0.299	0.381	0.676	0.872	1.237	1.635	2.204	2.661	3.070
7	0.485	0.598	0.989	1.239	1.690	2.167	2.833	3.358	3.822
8	0.710	0.857	1.344	1.646	2.180	2.733	3.490	4.078	4.594
9	0.972	1.152	1.735	2.088	2.700	3.325	4.168	4.817	5.380
10	1.265	1.479	2.156	2.558	3.247	3.940	4.865	5.570	6.179
11	1.587	1.834	2.603	3.053	3.816	4.575	5.578	6.336	6.989
12	1.934	2.214	3.074	3.571	4.404	5.226	6.304	7.114	7.807
13	2.305	2.617	3.565	4.107	5.009	5.892	7.042	7.901	8.634

Comparing two populations

	C4.5	C4.5+m	difference	rank
adult (sample)	0.763	0.768	+0.005	3.5
breast cancer	0.599	0.591	−0.008	7
breast cancer wisconsin	0.954	0.971	+0.017	9
cmc	0.628	0.661	+0.033	12
ionosphere	0.882	0.888	+0.006	5
iris	0.936	0.931	−0.005	3.5
liver disorders	0.661	0.668	+0.007	6
lung cancer	0.583	0.583	0.000	1.5
lymphography	0.775	0.838	+0.063	14
mushroom	1.000	1.000	0.000	1.5
primary tumor	0.940	0.962	+0.022	11
rheum	0.619	0.666	+0.047	13
voting	0.972	0.981	+0.009	8
wine	0.957	0.978	+0.021	10

[Janez Demsăr. From Statistical Comparisons of Classifiers over Multiple Data Sets. Journal of Machine Learning Research 7:1–30, 2006]

Paired sample comparison

Assume we have a single population $X_1, \ldots, X_n$ observed through one variable but measured at two time instants, e.g., individual i before treatment (A_i) and after treatment (P_i). These are called *paired* samples.

We want to test whether there is a statistical difference between the A_i's and the P_i's or not. E.g., does a treatment/algorithm has some effect?

- H_0 = no effect $\Rightarrow E[A] = E[P]$
- H_1 = effect $\Rightarrow E[A] \neq E[P], E[A] > E[P], E[A] < E[P]$

Let's consider $Z_i = A_i - P_i$, which are n i.i.d. variables for which we have $E[Z] = E[A] - E[P]$.

Under H_0, we expect to have $E[Z] = 0$. So the test boils down to a fitting test of the mean of Z_i with the theoretical value 0.

Testing moment equality

Are two independent samples of resp. sizes n_1 and n_2 coming from the same population/distribution?

Gaussian case: $X_1 \rightsquigarrow \mathcal{N}(m_1, \sigma_1)$ and $X_2 \rightsquigarrow \mathcal{N}(m_2, \sigma_2)$

- test variance equality (unknown mean) $\Rightarrow$ Fisher-Snedecor

$$F_{n_1-1,n_2-1} = \left(\frac{n1S_1^2}{n1-1}\right)\left(\frac{n2S_2^2}{n2-1}\right)^{-1}$$

- test mean equality (equal variance) $\Rightarrow$ Student

$$T_{n_1+n_2-2} = \sqrt{n_1+n_2-2}\,\frac{(\overline{X}_1 - m_1) - (\overline{X}_2 - m_2)}{\sqrt{(n_1S_1^2 + n_2S_2^2)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

UNIVERSITÉ DE RENNES 1

Wilcoxon tests

Are two independent samples of resp. sizes n_1 and n_2 coming from the same population/distribution?

General case: compare $\{x_1, \ldots, x_n\}$ and $\{y_1, \ldots, y_m\}$

- Wilcoxon
  - mix all values x_i and y_i and sort them (ascending order)
  - test statistics U = number of pairs (x_i, y_j) such that $x_i > y_j$
    - $U = 0 \Rightarrow x_1, \ldots, x_n, y_1, \ldots y_m$
    - $U = nm \Rightarrow y_1, \ldots y_m, x_1, \ldots, x_n$
    - under H_0 the ranking should be "homogeneous"
  - asymptotic distribution is $\mathcal{N}\left(\frac{nm}{2}, \sqrt{\frac{nm(n+m+1)}{12}}\right)$
  - critical region $\left|U - \frac{nm}{2}\right| > k_\alpha$

UNIVERSITÉ DE RENNES 1

The χ^2 test: are all subsamples similar?

TABLEAU 15.7.

	Modalité 1	Modalité 2		Modalité r	Total
Échantillon 1	n_{11}	n_{12}		n_{1r}	$n_{1.}$
Échantillon 2	n_{21}	n_{22}		n_{2r}	$n_{2.}$
⋮					
Échantillon k	n_{k1}	n_{k2}		n_{kr}	$n_{k.}$
Total	$n_{.1}$	$n_{.2}$		$n_{.r}$	n

- Test statistic

$$D^2 = \sum_i \sum_j \frac{(n_{ij} - n_{i.} p_j)^2}{n_{i.} p_j} = \sum_i \sum_j \frac{\left(n_{ij} - \frac{n_{i.} n_{.j}}{n}\right)^2}{\frac{n_{i.} n_{.j}}{n}}$$

- Under H_0 = all sub-samples have the same distribution

$$D^2 \rightsquigarrow \chi^2_{(k-1)(r-1)}$$

UNIVERSITÉ DE RENNES 1

Optimal decision for simple tests

Consider X with density $f(x;\theta)$, and denote $L(\mathbf{x},\theta)$ the density of the sample $\mathbf{x}$.

$$\begin{array}{ll} H_0 & \theta = \theta_0 \\ H_1 & \theta = \theta_1 \end{array}$$

Maximize the power of the test!

$$\Downarrow$$

$$P[W|H_1] = 1 - \beta = \int_W L(\mathbf{x};\theta_1)dx$$

Jerzy Neyman, 1894 – 1981

Karl Pearson, 1857 – 1936

The optimal critical region is defined by the points such that $\dfrac{L(\mathbf{x};\theta_1)}{L(\mathbf{x};\theta_0)} > k_\alpha$.

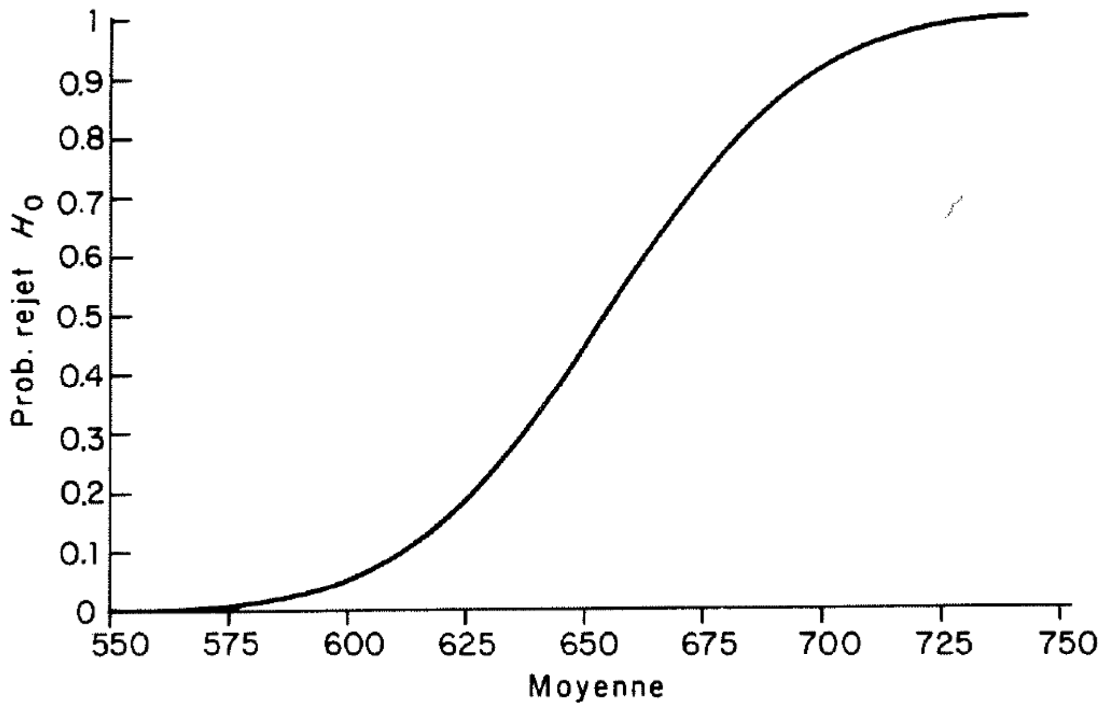
UNIVERSITÉ DE RENNES 1

Testing composite hypotheses

$$\begin{array}{ll} H_0 & \theta = \theta_0 \\ H_1 & \theta = \theta_1 \end{array} \qquad \text{vs} \qquad \begin{array}{ll} H_0 & \theta = \theta_0 \\ H_1 & \theta \neq \theta_0 \end{array}$$

$\Rightarrow$ the risk β (and hence the power) depends on θ

$$\begin{array}{ll} H_0 & m = 600 \\ H_1 & m > 600 \end{array}$$



The likelihood ratio test

$$\begin{array}{ll} H_0 & \theta = \theta_0 \\ H_1 & \theta \neq \theta_0 \end{array} \qquad \text{with } \theta \in \mathbb{R}^p$$

- Test statistics

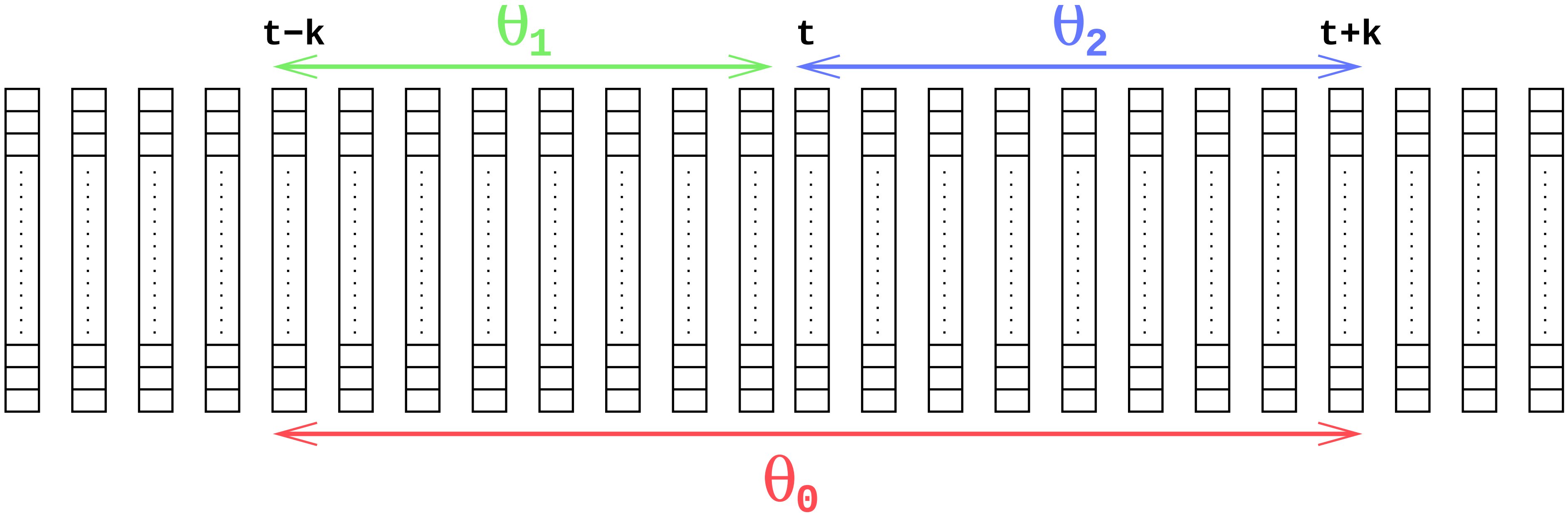
$$\lambda = \frac{L(\mathbf{x};\theta_0)}{\sup\limits_{\theta} L(\mathbf{x};\theta)}$$

  - Note that replacing $L(\mathbf{x};\theta_0)$ by $\sup\limits_{\theta} L(\mathbf{x};\theta)$ is like using the ML estimate of θ

- Critical region = $\{x|\lambda < K_\alpha\}$
- Asymptotic distribution of the test statistics under H_0: $-2\ln(\lambda) \rightsquigarrow \chi^2_p$

UNIVERSITÉ DE RENNES 1

Detecting changes in statistics



Speaker or face identity verification

- H_0: the person is who he/she says he/she is
- H_1: the person is an impostor

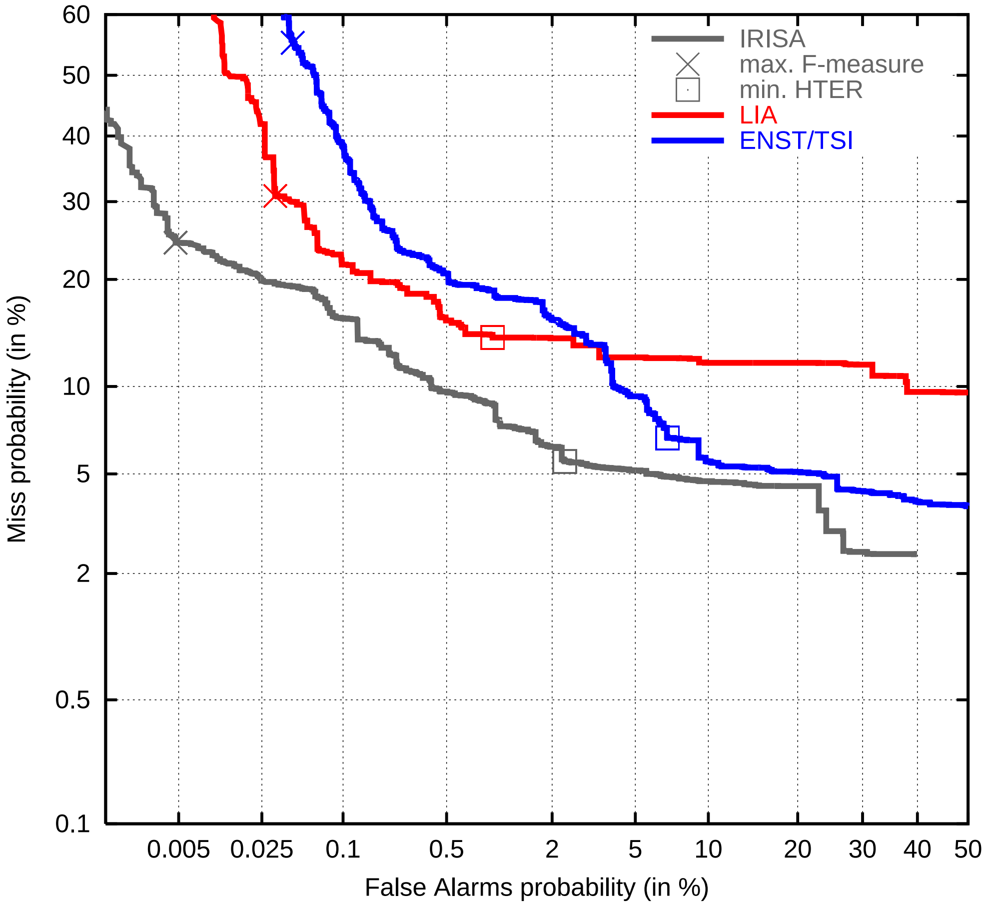
$$\frac{p(y_1^T; H_0)}{p(y_1^T; H_1)} \underset{H_1}{\overset{H_0}{\gtrless}} \beta$$

In speaker verification

- $p(y_1^T; H_0)$ = Gaussian mixture model trained with the speaker's speech
- $p(y_1^T; H_1)$ = GMM of speech in general

Speaker or face identity verification (cont'd)

Two errors : false acceptance (type 1) et false rejection (type 2).



- text known or not
- size of the data set
- signal duration
- signal quality
- speaker
- ...

site	ENST	IRISA	LIA
%fa	9.8	0.3	2.8
%fr	25.3	23.6	30.6
F-measure	46.9	84.3	66.0

Thanks for attending until the end!