
METODI STATISTICI E COMPUTAZIONALI

Stefania Spagnolo

Dipartimento di Matematica e Fisica, Univ. del Salento



LEZIONE 6

RICHIAMI DI CALCOLO COMBINATORIO E PROBABILITÀ

Continua

CALCOLO COMBINATORIO E PROBABILITÀ

- Esempio:
- Supponiamo di avere a che fare con una malattia rara che colpisce la popolazione con una frequenza di 2 malati ogni 100000 abitanti.
- $P(\text{malato}) = 2/100000 = 2 \times 10^{-5}$

NOTA: probabilità nota a priori
- Supponiamo di disporre di un test con le seguenti caratteristiche:
- **Sensibilità:** probabilità, cioè, di un risultato positivo se la persona è effettivamente malata

Calibrazione. Conosco quale è l'effetto data la causa.

 - $P(\text{pos} \mid \text{malato}) = 0.993$
- **Specificità:** probabilità, cioè, di un risultato negativo se la persona è sana
 - $P(\text{neg} \mid \text{sano}) = 0.9999$

Il test è molto buono. La Sensibilità mi sta dicendo che identifica un malato se questo lo è effettivamente con una probabilità del 99.3 % mentre la specificità mi dice che dichiara una persona sana se questa lo è effettivamente con una probabilità del 99.99%

CALCOLO COMBINATORIO E PROBABILITÀ

$$P(B_k | A_i) = \frac{P(A_i | B_k) P(B_k)}{\sum_j P(A_i | B_j) P(B_j)}$$

Notiamo come sia la specificità che la sensibilità del test sono due esempi di probabilità condizionata.

Misura

Ora supponiamo di voler applicare il test ad un campione di persone a caso la domanda che ci poniamo è:

“ quanti di coloro che sono risultati positivi al test sono effettivamente malati e quanti invece sono sani? ”

Questo equivale ad identificare la probabilità (condizionata) che un soggetto positivo al test sia effettivamente malato

Questa è conoscenza **a priori**

$$P(\text{malato} | \text{pos}) = \frac{P(\text{pos} | \text{malato})P(\text{malato})}{P(\text{pos} | \text{malato})P(\text{malato}) + P(\text{pos} | \text{sano})P(\text{sano})}$$

CALCOLO COMBINATORIO E PROBABILITÀ

$$P(\text{malato} | \text{pos}) = \frac{P(\text{pos} | \text{malato})P(\text{malato})}{P(\text{pos} | \text{malato})P(\text{malato}) + P(\text{pos} | \text{sano})P(\text{sano})}$$

$$P(\text{pos} | \text{malato}) = 0.993$$

Sensibilità del test

$$P(\text{malato}) = 2 \cdot 10^{-5}$$

Conoscenza a priori

$$P(\text{pos} | \text{sano}) = 1 - P(\text{neg} | \text{sano}) = 1 - 0.9999 = 1 \cdot 10^{-4}$$

Da specificità

$$P(\text{sano}) = 1 - P(\text{malato}) = 0.99998$$

Conoscenza a priori

Da cui

$$P(\text{malato} | \text{pos}) = 0.1656$$

Per cui di coloro che risultano positivi al test solo poco più del 16% sono effettivamente malati i rimanenti (84%) sono sani ma erroneamente indicati come malati (falsi positivi)

CALCOLO COMBINATORIO E PROBABILITÀ

Quant'è invece la frazione di coloro che risultano negativi al test e sono effettivamente sani

Questo equivale a calcolare la $P(\text{sano}|\text{neg})$

$$P(\text{sano} | \text{neg}) = \frac{P(\text{neg} | \text{sano})P(\text{sano})}{P(\text{neg} | \text{sano})P(\text{sano}) + P(\text{neg} | \text{malato})P(\text{malato})}$$

Con un calcolo analogo al precedente si trova che

$$P(\text{sano} | \text{neg}) = 0.99999999$$

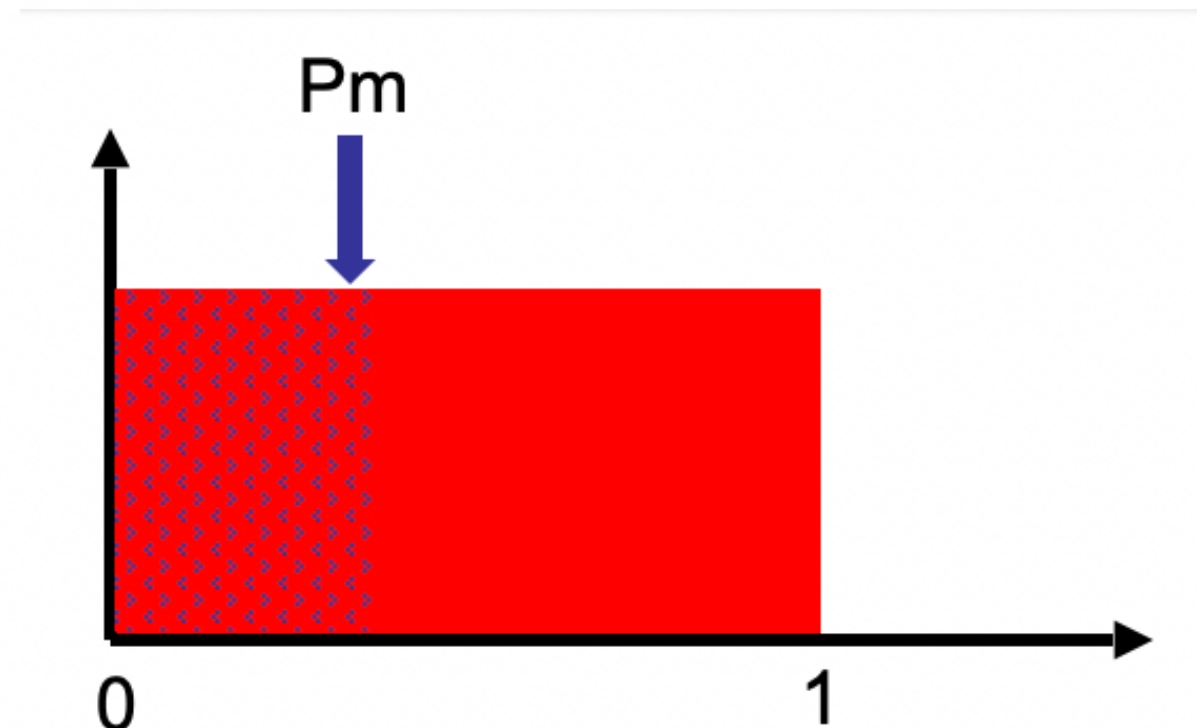
Quindi se il risultato del test è negativo, la probabilità di essere sani è estremamente elevata o, inversamente, la probabilità che un malato risulti negativo al test (falso negativo) è estremamente bassa.

SIMULAZIONE

```
#include "TRandom.h"
void Bayes(long N=1000000){
    double Pm= 0.00002; // a priori
    //double Pm= 0.001; // a priori
    double Ppm=0.993; // P(positivo|malato)
    double Pns=0.9999; // P(negativo|sano)
    TRandom3 R(12345);
    long Pos=0; // contatore esiti positivi
    long Neg=0; // contatore esiti positivi
    long MVer0=0; // contatore malati (veri)

    //long N=10000000000;
    //long N=100000000;
    long FalsiNegativi=0;
    long VeriNegativi=0;
    long FalsiPositivi=0;
    long VeriPositivi=0;

    for (int i=0; i<N ; i++) {
        int Sog=0; // =0 -> sano ; =1 -> malato
        double x=R.Uniform();
        int R1=0;
        if (x<Pm) {
            Sog=1; // malato
            MVer0++;
        }
    }
}
```

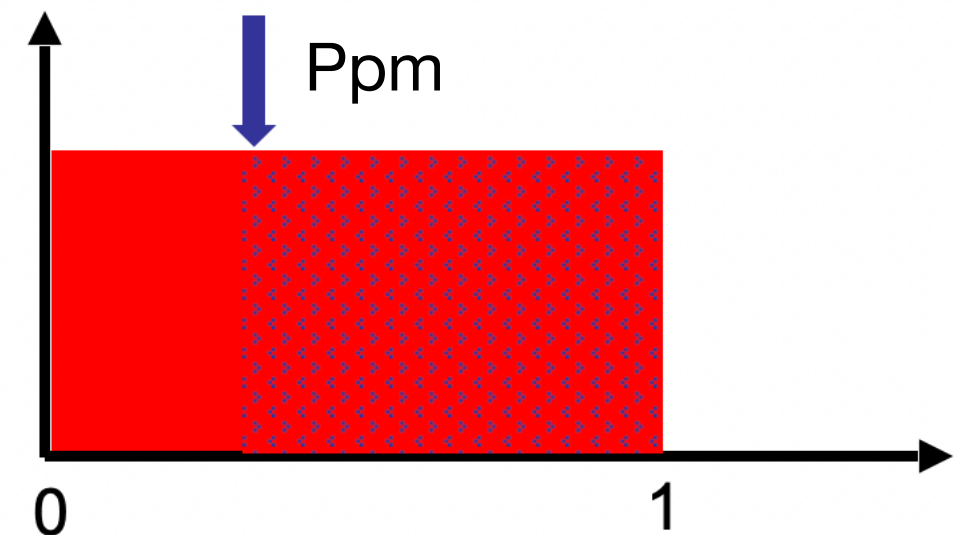
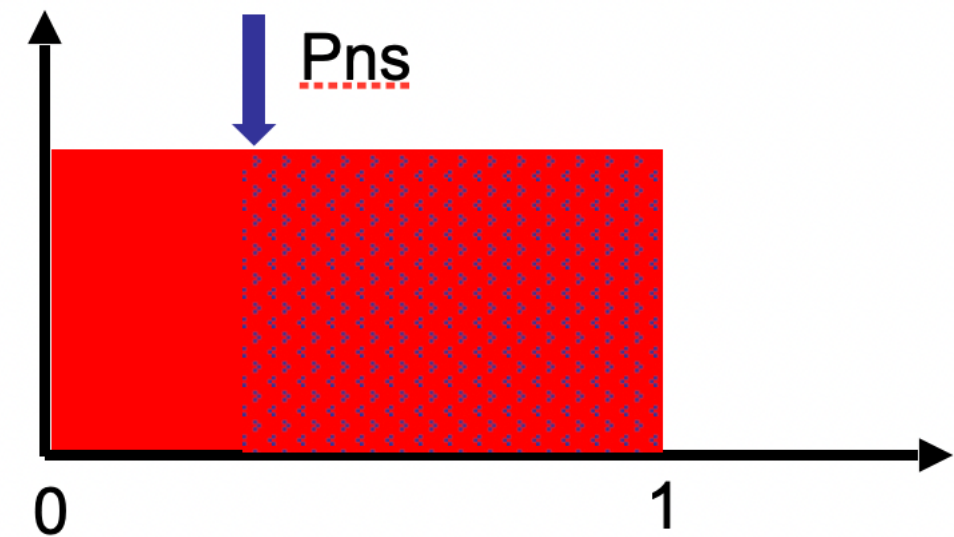


Sog=1 → Malato

Sog=0 → Sano

SIMULAZIONE

```
x=R.Uniform();
if (Sog==0)
{
    // se soggetto sano
    if (x<Pns) {
        // esito negativo per soggetto sano
        R1=1;
        VeriNegativi++;
        Neg++; // esisto negativo
    }
    else
    {
        Pos++; // esisto positivo
        FalsiPositivi++;
    }
}
else
{
    // se soggetto malato
    if (x<Ppm) {
        // esito positivo per soggetto malato
        R1=1;
        Pos++; // esisto positivo
        VeriPositivi++;
    }
    else
    {
        Neg++; // esisto negativo
        FalsiNegativi++;
    }
}
```



CLASSIFICAZIONE

- Metrica di un classificatore
- Curva ROC

TRUE / FALSE (TRUTH LABEL) & POSITIVE / NEGATIVE (TEST LABEL)

A **true positive** is an outcome where the model *correctly* predicts the *positive* class. Similarly, a **true negative** is an outcome where the model *correctly* predicts the *negative* class.

A **false positive** is an outcome where the model *incorrectly* predicts the *positive* class. And a **false negative** is an outcome where the model *incorrectly* predicts the *negative* class.

TP = signal event labelled as signal, i.e. giving a positive test result

TN = background event labelled as background, i.e. giving a negative test result

FP = background event labelled as signal, i.e. with positive test

FN = signal event labelled as background, i.e. with negative test

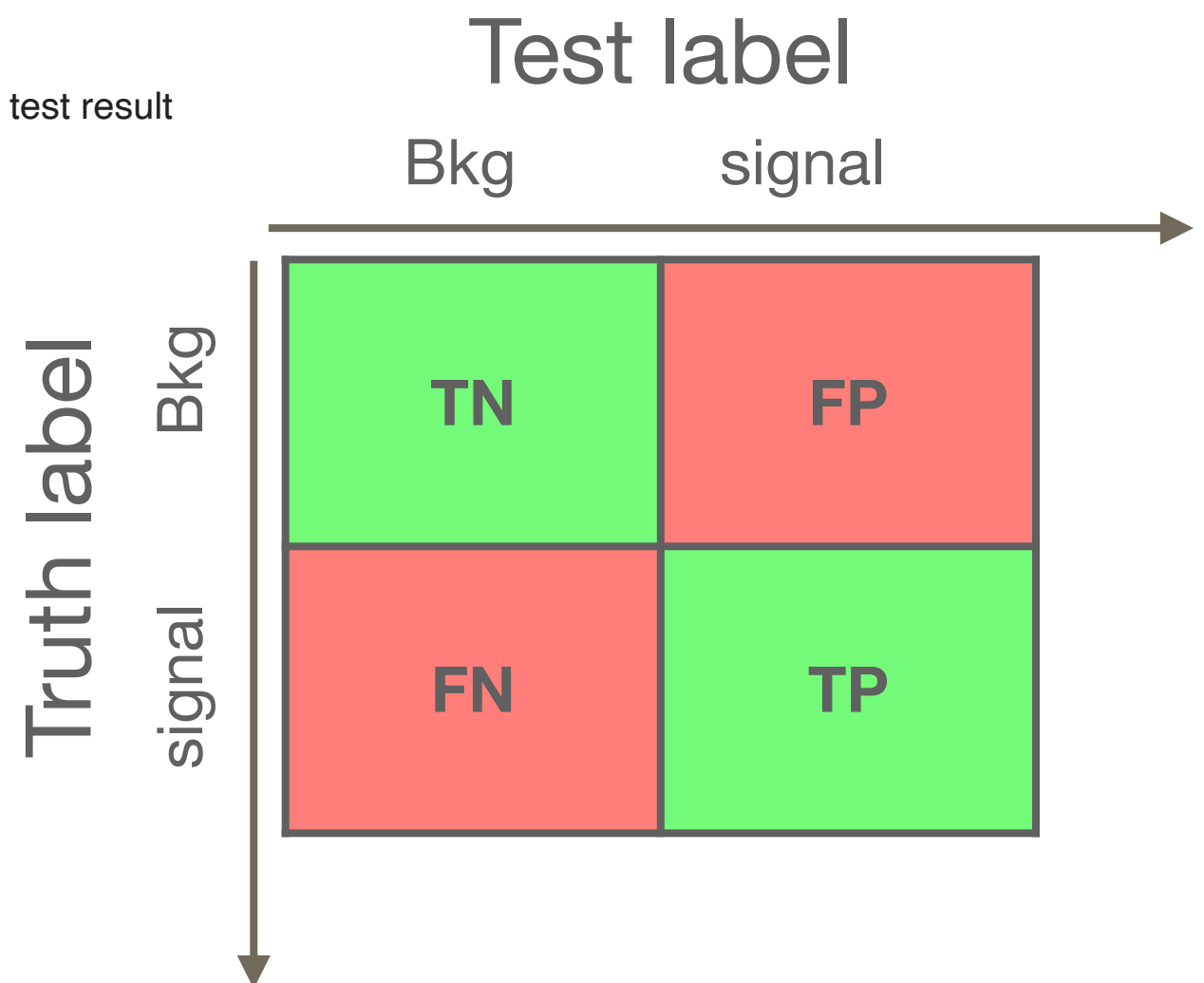
Signal events = TP + FN

Background events = FP + TN

Signal efficiency: $\frac{TP}{TP + FN}$

Background efficiency: $\frac{FP}{FP + TN}$

Background rejection: $\frac{FP + TN}{FP} = 1 + \frac{TN}{FP} \simeq \frac{TN}{FP}$



Our confusion matrix definition

TRUE / FALSE & POSITIVE / NEGATIVE

Other definitions:

False positive rate = def = $1 - \text{specificity}$

True positive rate = def = sensitivity

$\text{Sensitivity} = \frac{TP}{TP + FN}$ ==>> fraction of ill people correctly identified

Sensitivity = Signal efficiency

$\text{Specificity} = \frac{TN}{TN + FP}$ ==>> fraction of healthy people correctly recognised as such

1 - specificity = Background efficiency

Hence

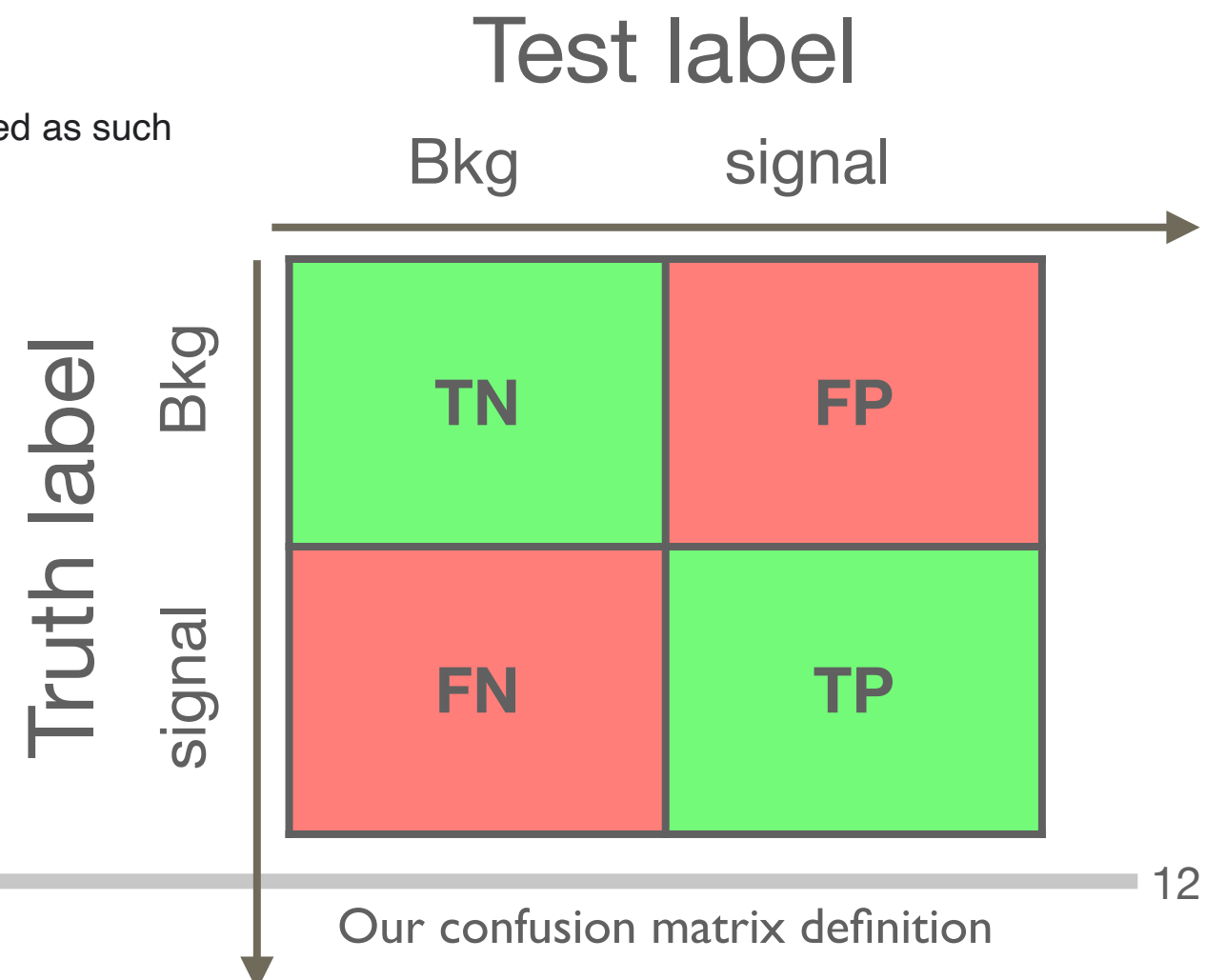
FPrate = background efficiency

TPrate = signal efficiency

ROC = TPrate vs FPrate

= signal eff. vs background off.

$$\text{FPrate} = 1 - \text{specificity} = 1 - \frac{TN}{TN + FP} = \frac{FP}{TN + FP}$$
$$\text{TPrate} = \text{sensitivity} = \frac{TP}{TP + FN}$$



ACCURACY

$$\text{Accuracy} = \text{Number of correct predictions} / \text{total number of predictions} = \frac{TP + TN}{TP + TN + FP + FN}$$

Accuracy is not a good metric for a class-imbalanced data set +++++++>>>>>>>>
*when quoting **ACCURACY** we **MUST** balance the data set through **weights** !*

Example

Let's consider a tumor-classifier model with a given sensitivity and specificity and a screening sample of 100 people,

91 healthy -> for 1/ 91 the test gives a positive answer

9 sick -> for 1/ 9 the test gives a positive answer

ACCURACY

$$\text{Accuracy} = \text{Number of correct predictions} / \text{total number of predictions} = \frac{TP + TN}{TP + TN + FP + FN}$$

Accuracy is not a good metric for a class-imbalanced data set +++++++>>>>>>>>
*when quoting **ACCURACY** we MUST balance the data set through **weights** !*

The diagram shows a horizontal axis with a green arrow pointing right, labeled 's' at the start and 'b' at the end. Above the arrow, the text 'Test label' is written. Below the arrow, four colored boxes represent different classification outcomes:

- True Positive (TP):** A green box containing:
 - Reality: Malignant
 - ML model predicted: Malignant
 - Number of TP results: 1
- False Negative (FN):** A red box containing:
 - Reality: Malignant
 - ML model predicted: Benign
 - Number of FN results: 8
- False Positive (FP):** A red box containing:
 - Reality: Benign
 - ML model predicted: Malignant
 - Number of FP results: 1
- True Negative (TN):** A green box containing:
 - Reality: Benign
 - ML model predicted: Benign
 - Number of TN results: 90

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{1 + 90}{1 + 90 + 1 + 8} = 0.91$$

BUT !!!! 8 out of 9 malignancies go undiagnosed

Another tumor-classifier model that always predicts benign would achieve the exact same accuracy (91/100 correct predictions) on our example

ACCURACY

$$\text{Accuracy} = \text{Number of correct predictions} / \text{total number of predictions} = \frac{TP + TN}{TP + TN + FP + FN}$$

Accuracy is not a good metric for a class-imbalanced data set +++++++>>>>>>>>>
*when quoting **ACCURACY** we **MUST** balance the data set through **weights** !*

Example

Let's consider a tumor-classifier model with a given sensitivity and specificity and a screening sample of 100 people,

91 healthy -> for 1/ 91 the test gives a positive answer

9 sick -> for 1/ 9 the test gives a positive answer

Extreme example: a tumor-classifier model that always predicts benign (sensitivity = 0, specificity 100%)

On the same screening sample of 100 people,

91 healthy -> for 0/ 91 the test gives a positive answer

9 sick -> for 0/ 9 the test gives a positive answer

ACCURACY

$$\text{Accuracy} = \text{Number of correct predictions} / \text{total number of predictions} = \frac{TP + TN}{TP + TN + FP + FN}$$

Accuracy is not a good metric for a class-imbalanced data set +++++++>>>>>>>
when quoting **ACCURACY** we MUST balance the data set through **weights** !

The diagram shows a horizontal axis with 's' (signal) on the left and 'b' (background) on the right, separated by a vertical line. An arrow labeled 'Test label' points to the right. The results are categorized into four quadrants:

- True Positive (TP):** Reality: Malignant, ML model predicted: Malignant. Truth label: signal, Test label: signal. Number of TP results: 0.
- False Negative (FN):** Reality: Malignant, ML model predicted: Benign. Truth label: signal, Test label: bkg. Number of FN results: 9.
- False Positive (FP):** Reality: Benign, ML model predicted: Malignant. Truth label: bkg, Test label: signal. Number of FP results: 0.
- True Negative (TN):** Reality: Benign, ML model predicted: Benign. Truth label: bkg, Test label: bkg. Number of TN results: 91.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{0 + 91}{0 + 91 + 0 + 9} = 0.91$$

100% of malignancies go undiagnosed

Another tumor-classifier model that always predicts benign would achieve the exact same accuracy (91/100 correct predictions) on our example

PRECISION AND RECALL

Back to our tumor-classifier model
 screening sample of 100 people,
 91 healthy -> for 1/ 91 the test gives a positive answer
 9 sick -> for 1/ 9 the test gives a positive answer

$$\text{Precision} = \frac{TP}{TP + FP}$$

What is the fraction of positive tests that is correct ?

Precision is 1 if the model produces no FP

Precision = 0.5
 Recall = 1 / 9 = 0.11

		Test label →	
Truth label ↓	signal	True Positive (TP): - Reality: Malignant - ML model predicted: Malignant - Number of TP results: 1	False Negative (FN): - Reality: Malignant - ML model predicted: Benign - Number of FN results: 8
	bkg	False Positive (FP): - Reality: Benign - ML model predicted: Malignant - Number of FP results: 1	True Negative (TN): - Reality: Benign - ML model predicted: Benign - Number of TN results: 90

$$\text{Recall} = \frac{TP}{TP + FN}$$

What is the fraction of true events that is correctly identified by the test ?

$$\text{Recall} = \text{signal efficiency} = \frac{TP}{\text{Signal events}}$$

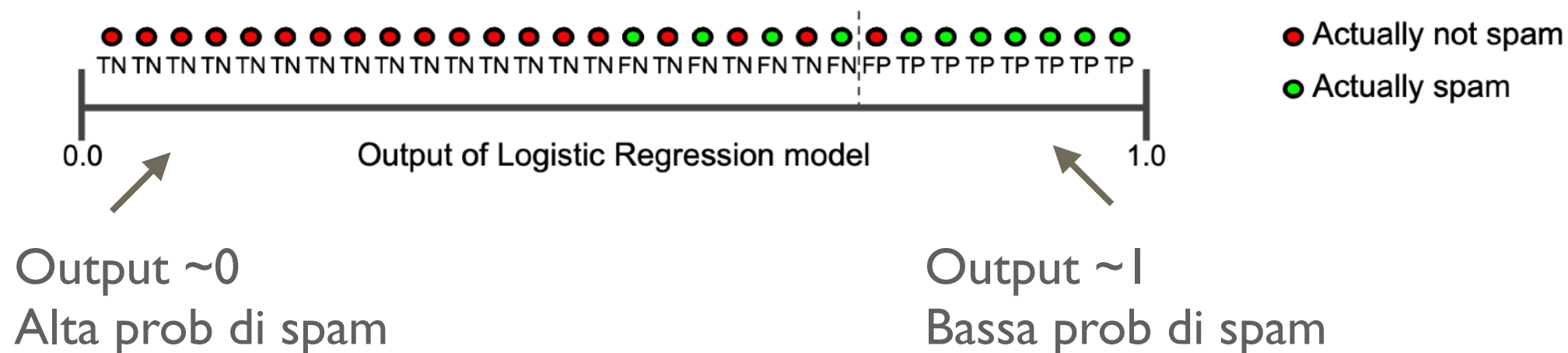
In a classification problem, the decision threshold define the Precision and Recall (efficiency):

- hard threshold means high precision, low efficiency

PRECISION, RECALL AND DECISION THRESHOLD

Consideriamo il problema di riconoscere i mail spam.

Immaginiamo di avere un algoritmo che dia in Output la probabilità che un e-mail sia / non sia spam



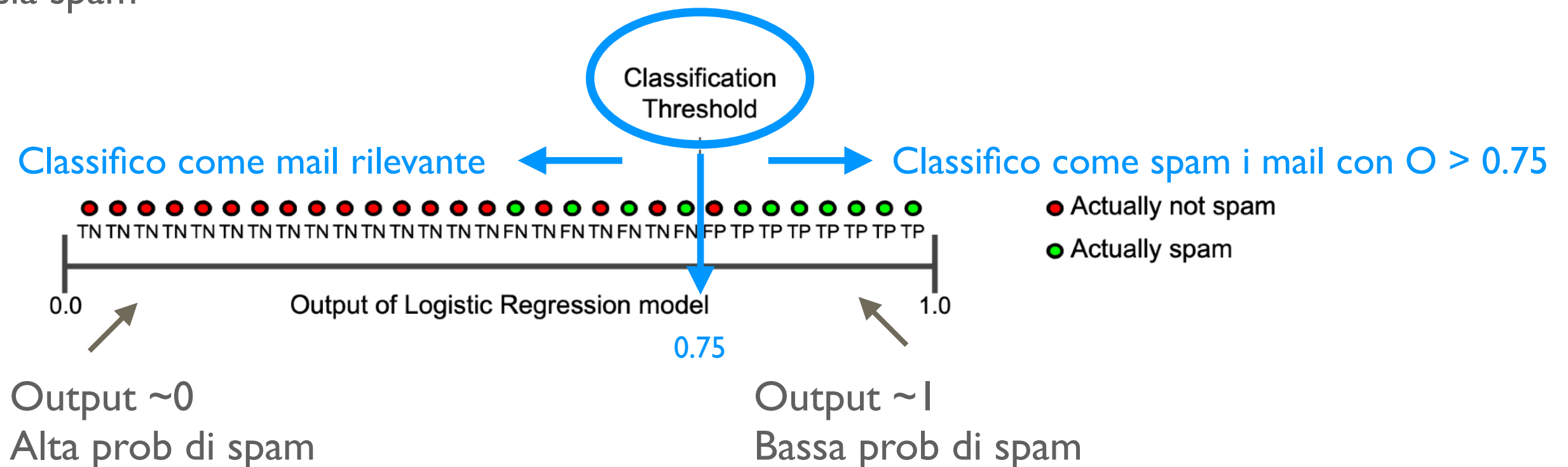
Per decidere quali mail cestinare, ordiniamoli in base al valore dell'Output dell'algoritmo;
il **colore verde** indica mail realmente **spam**
il **colore rosso** indica mail che **non** sono **spam**.

Ma il colore (l'etichetta di realtà) non è noto

PRECISION, RECALL AND DECISION THRESHOLD

Consideriamo il problema di riconoscere i mail spam.

Immaginiamo di avere un algoritmo che dia in Output la probabilità che un e-mail sia / non sia spam

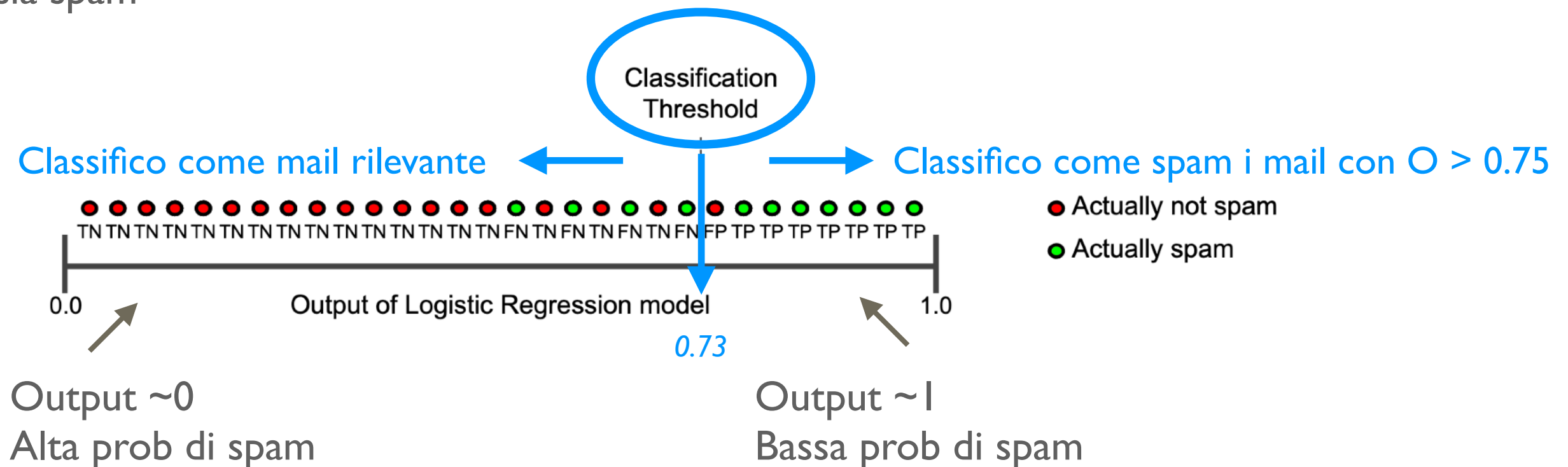


Per CLASSIFICARE come NON-spam un e-mail applico **una soglia** sull'output dell'algoritmo

PRECISION, RECALL AND DECISION THRESHOLD

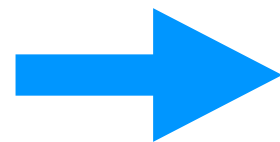
Consideriamo il problema di riconoscere i mail spam.

Immaginiamo di avere un algoritmo che dia in Output la probabilità che un e-mail sia / non sia spam



Per CLASSIFICARE come NON-spam un e-mail applico **una soglia** sull'output dell'algoritmo
La mia classificazione (la mia scelta della **soglia**) ha una certa **precision** e **recall**

Ad esempio, per
 $O_{thr} = 0.73$ ho
questa matrice di
confusione



True Positives (TP): 7	False Negatives (FN): 4
False Positives (FP): 1	True Negatives (TN): 18

PRECISION, RECALL AND DECISION THRESHOLD

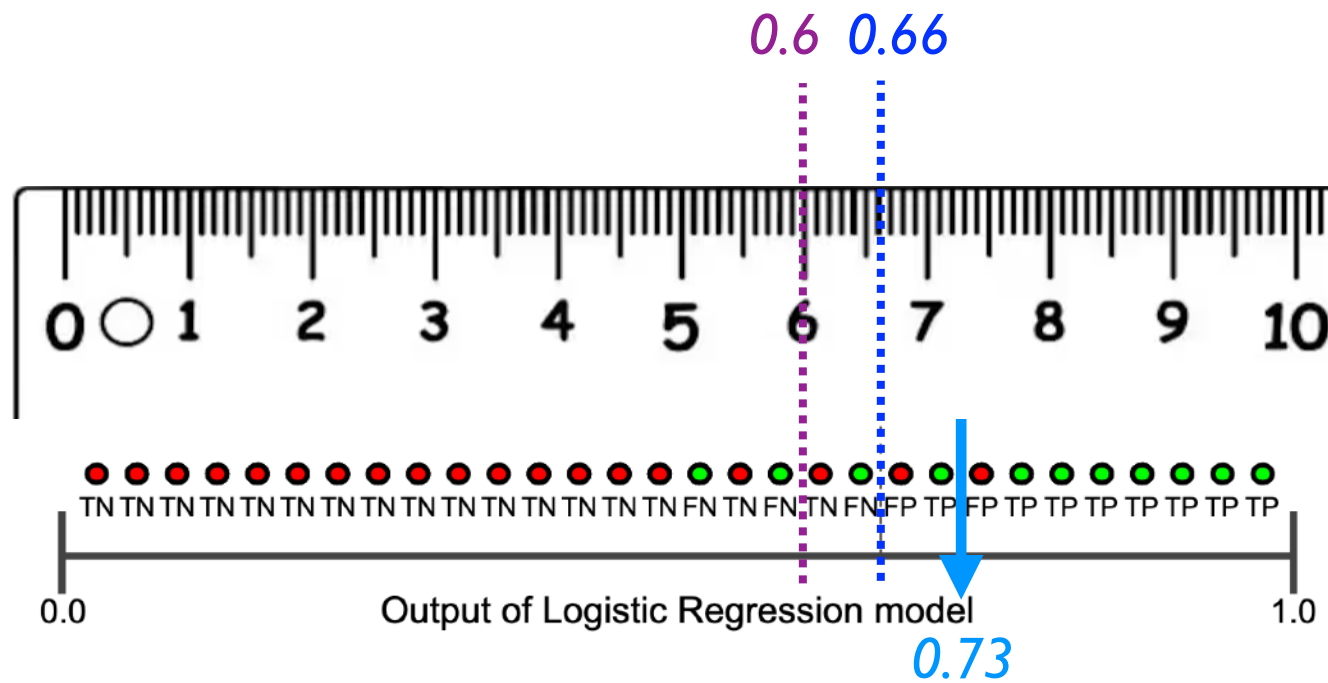


Figure 1. Classifying email messages as spam or not spam.

per $O_{thr} = 0.66$ la matrice di confusione diventa

True Positives (TP): 8	False Negative (FN): 3
False Positive (FP): 2	True Negatives (TN): 17

$$O_{thr} = 0.73$$

Precision = $7 / (7+1) = 0.88$
Recall = $7 / (7+4) = 0.64$

$$O_{thr} = 0.66$$

Precision = $8 / (8+2) = 0.8$
Recall = $8 / (8+3) = 0.73$

$$O_{thr} = 0.6$$

Precision = $9 / (9+3) = 0.75$
Recall = $9 / (9+2) = 0.82$

In a classification problem, the decision threshold define the Precision and Recall (efficiency):

- **hard threshold means high precision, low efficiency**

PRECISION, RECALL AND F_1 SCORE

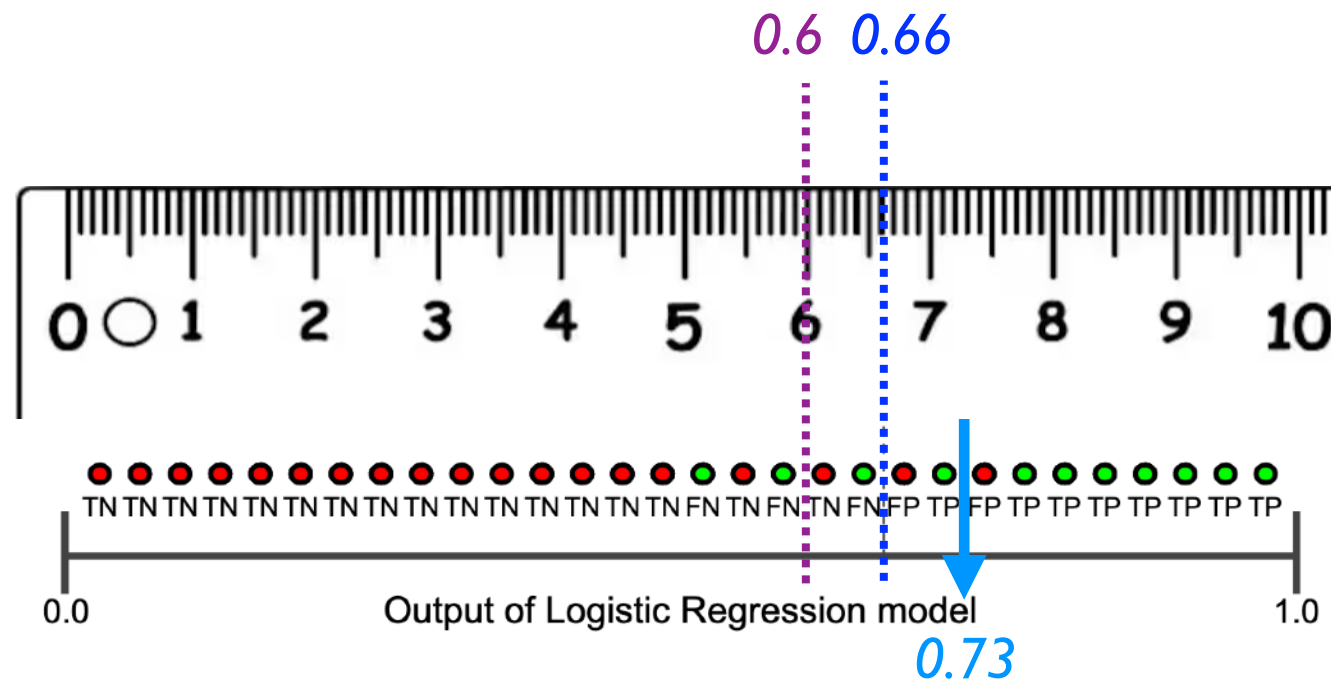


Figure 1. Classifying email messages as spam or not spam.

per $O_{thr} = 0.66$ la matrice di confusione diventa

True Positives (TP): 8	False Negative (FN): 3
False Positive (FP): 2	True Negatives (TN): 17

$$O_{thr} = 0.73$$

$$\text{Precision} = 7 / (7+1) = 0.88$$

$$\text{Recall} = 7 / (7+4) = 0.64$$

$$\text{FPrate} = 1 / 19 = 0.052$$

$$F_1\text{-score} = 0.74$$

$$O_{thr} = 0.66$$

$$\text{Precision} = 8 / (8+2) = 0.8$$

$$\text{Recall} = 8 / (8+3) = 0.73$$

$$\text{FPrate} = 2 / 19 = 0.105$$

$$F_1\text{-score} = 0.76$$

$$O_{thr} = 0.6$$

$$\text{Precision} = 9 / (9+3) = 0.75$$

$$\text{Recall} = 9 / (9+2) = 0.82$$

$$\text{FPrate} = 3 / 19 = 0.157$$

$$F_1\text{-score} = 0.78$$

F_1 -score = def = harmonic average of recall and precision

$$F_1 = \frac{2}{\text{recall}^{-1} + \text{precision}^{-1}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{2 \cdot \text{tp}}{2 \cdot \text{tp} + \text{fp} + \text{fn}}$$

THE CONFUSION MATRIX

Signal events = $TP + FN$ # Background events = $FP + TN$

Signal efficiency: $\frac{TP}{TP + FN}$ Sensitivity = $\frac{TP}{TP + FN}$

Background efficiency: $\frac{FP}{FP + TN}$

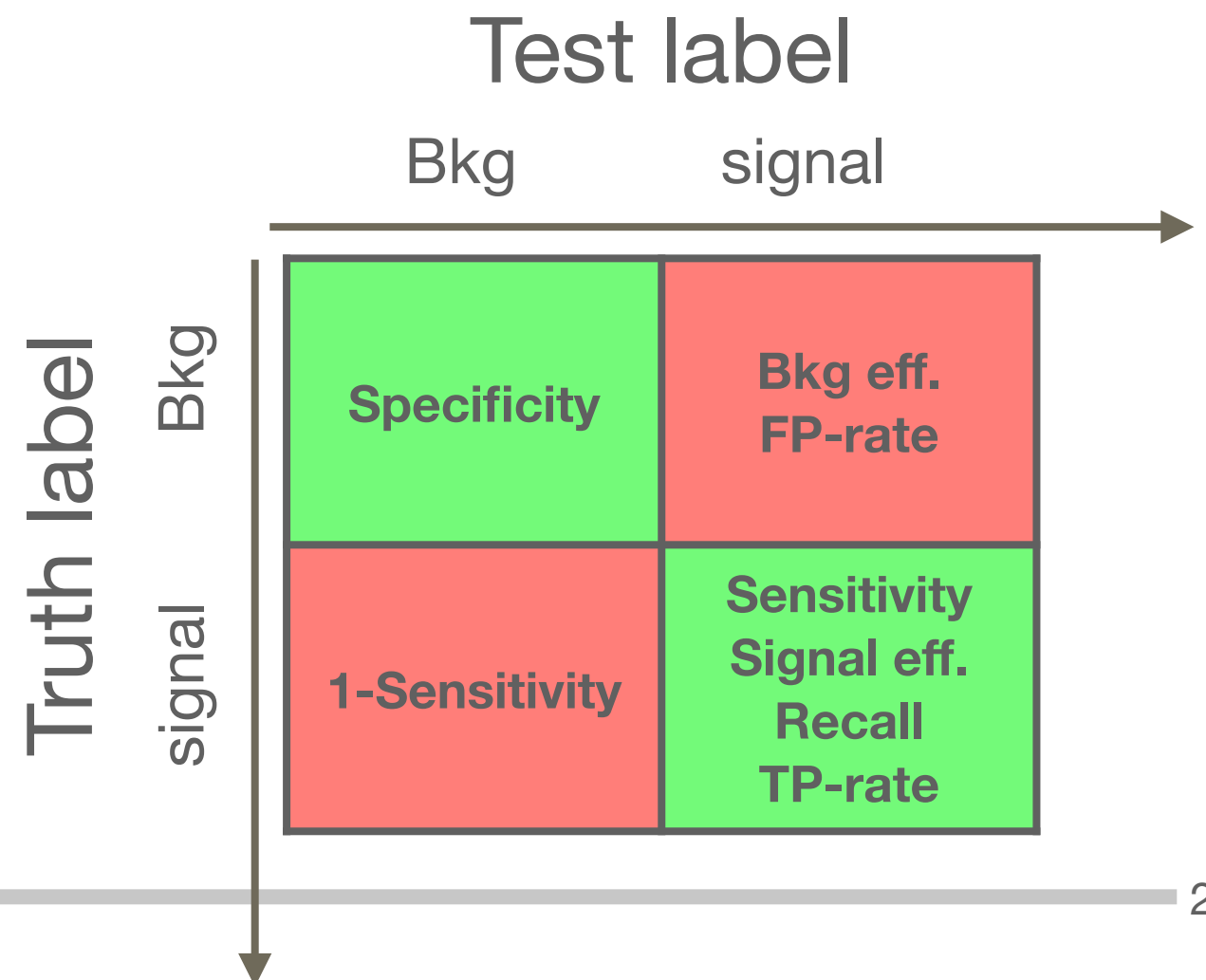
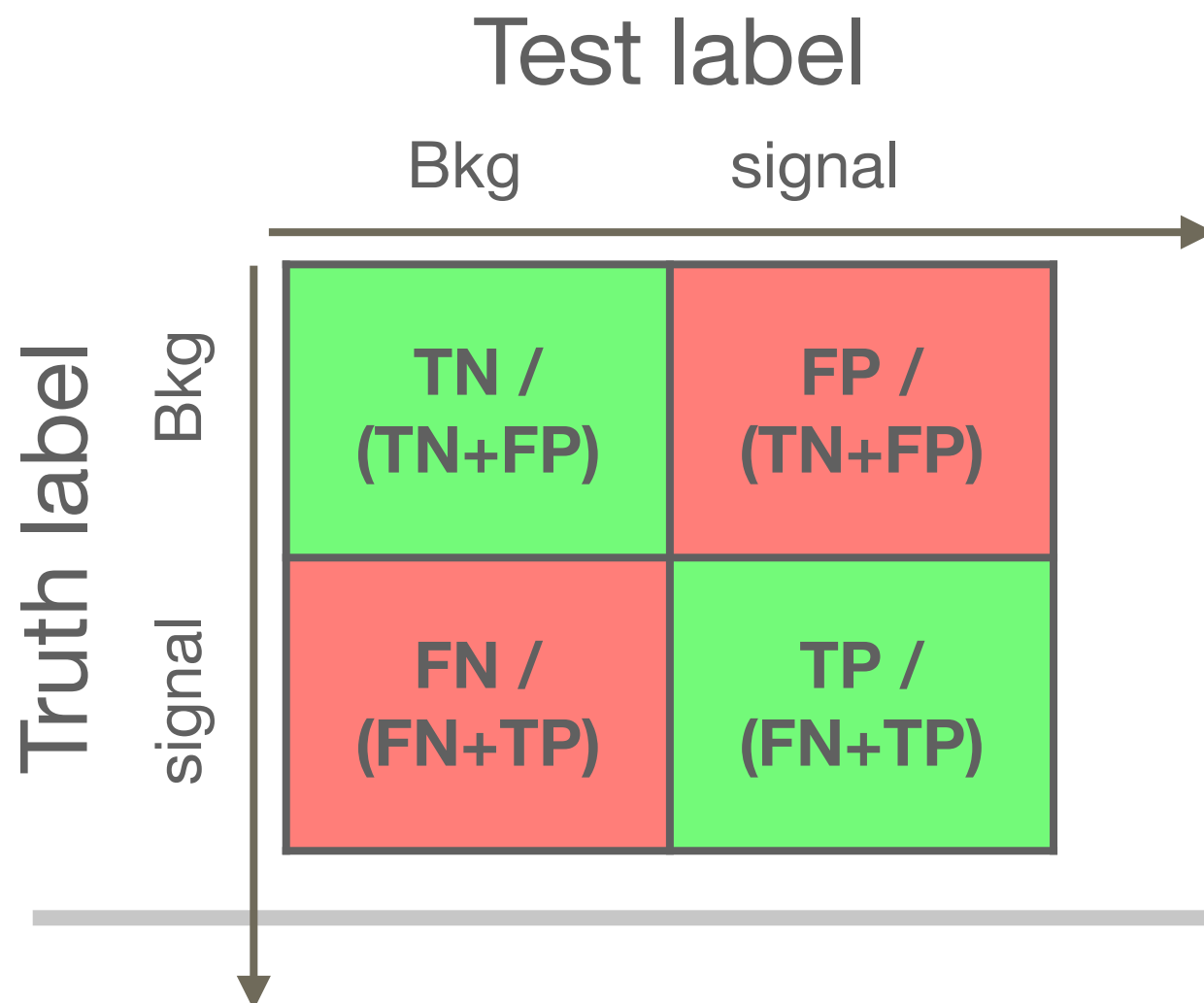
Specificity = $\frac{TN}{TN + FP}$

FPrate = 1 - specificity = $1 - \frac{TN}{TN + FP} = \frac{FP}{TN + FP}$

FPrate = background efficiency

TPrate = sensitivity = $\frac{TP}{TP + FN}$

TPrate = signal efficiency



THE ROC (RECEIVER OPERATING CHARACTERISTIC): TPrate vs FPrate

Signal events = $TP + FN$ # Background events = $FP + TN$

Signal efficiency: $\frac{TP}{TP + FN}$ Sensitivity = $\frac{TP}{TP + FN}$

Background efficiency: $\frac{FP}{FP + TN}$

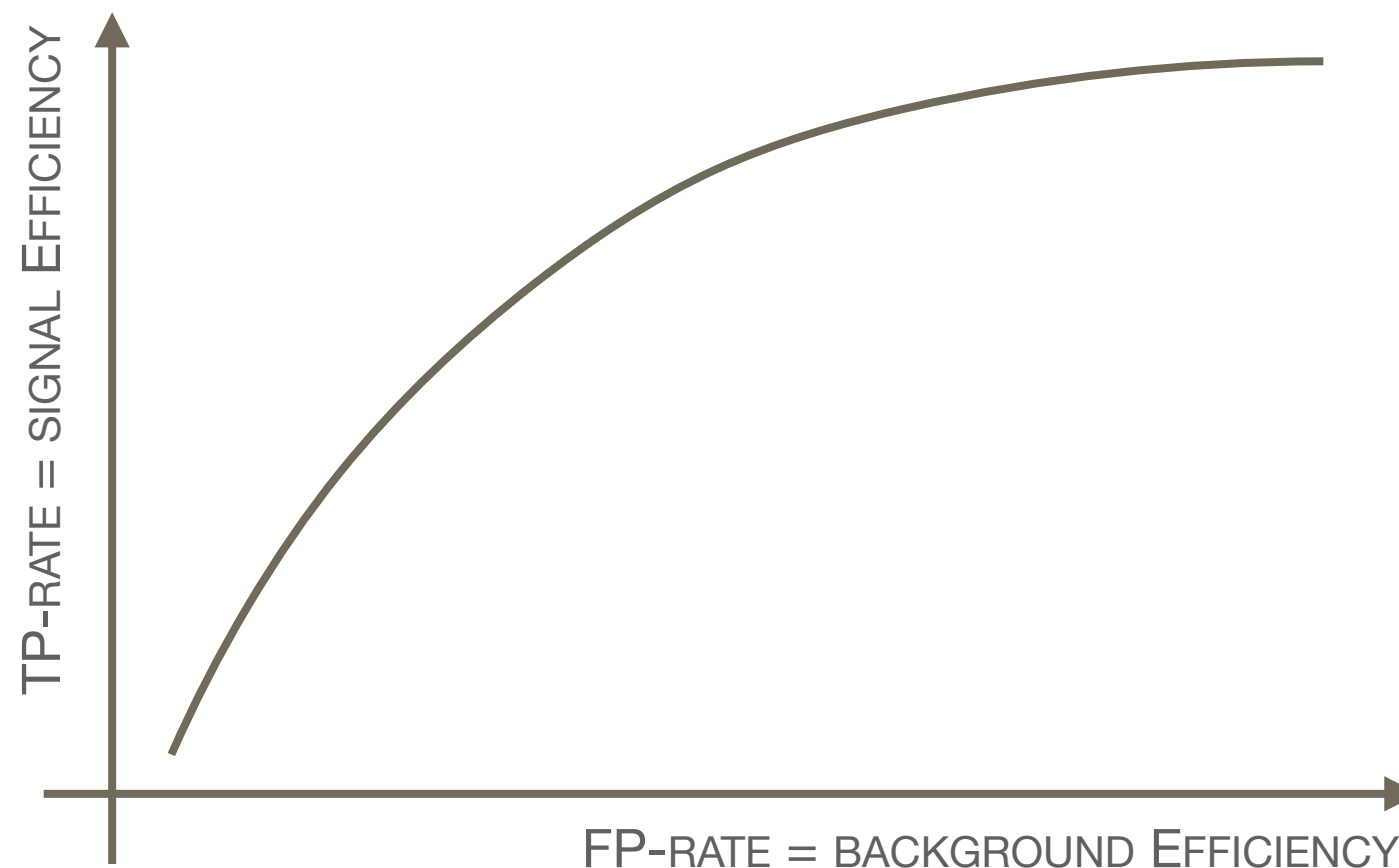
Specificity = $\frac{TN}{TN + FP}$

FPrate = 1 - specificity = $1 - \frac{TN}{TN + FP} = \frac{FP}{TN + FP}$

FPrate = background efficiency

TPrate = sensitivity = $\frac{TP}{TP + FN}$

TPrate = signal efficiency



THE ROC (RECEIVER OPERATING CHARACTERISTIC): TPRATE VS FPRATE

Signal events = TP + FN # Background events = FP + TN

Signal efficiency: $\frac{TP}{TP + FN}$ Sensitivity = $\frac{TP}{TP + FN}$

Background efficiency: $\frac{FP}{FP + TN}$

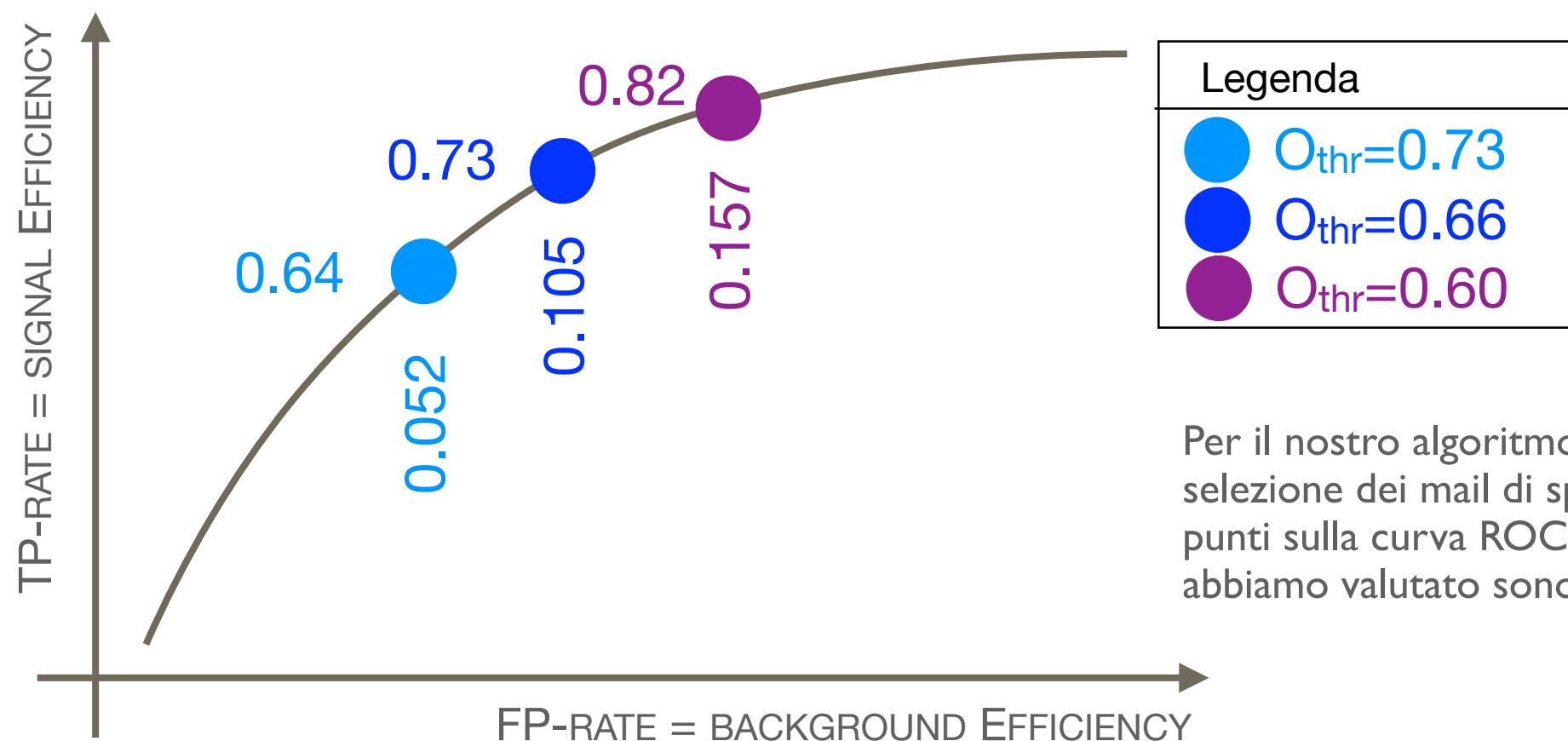
Specificity = $\frac{TN}{TN + FP}$

FPrate = 1 - specificity = $1 - \frac{TN}{TN + FP} = \frac{FP}{TN + FP}$

FPrate = background efficiency

TPrate = sensitivity = $\frac{TP}{TP + FN}$

TPrate = signal efficiency



Per il nostro algoritmo di selezione dei mail di spam i tre punti sulla curva ROC che abbiamo valutato sono in figura

THE ROC (RECEIVER OPERATING CHARACTERISTIC): TP RATE VS FP RATE

Signal events = $TP + FN$ # Background events = $FP + TN$

Signal efficiency: $\frac{TP}{TP + FN}$ Sensitivity = $\frac{TP}{TP + FN}$

Background efficiency: $\frac{FP}{FP + TN}$

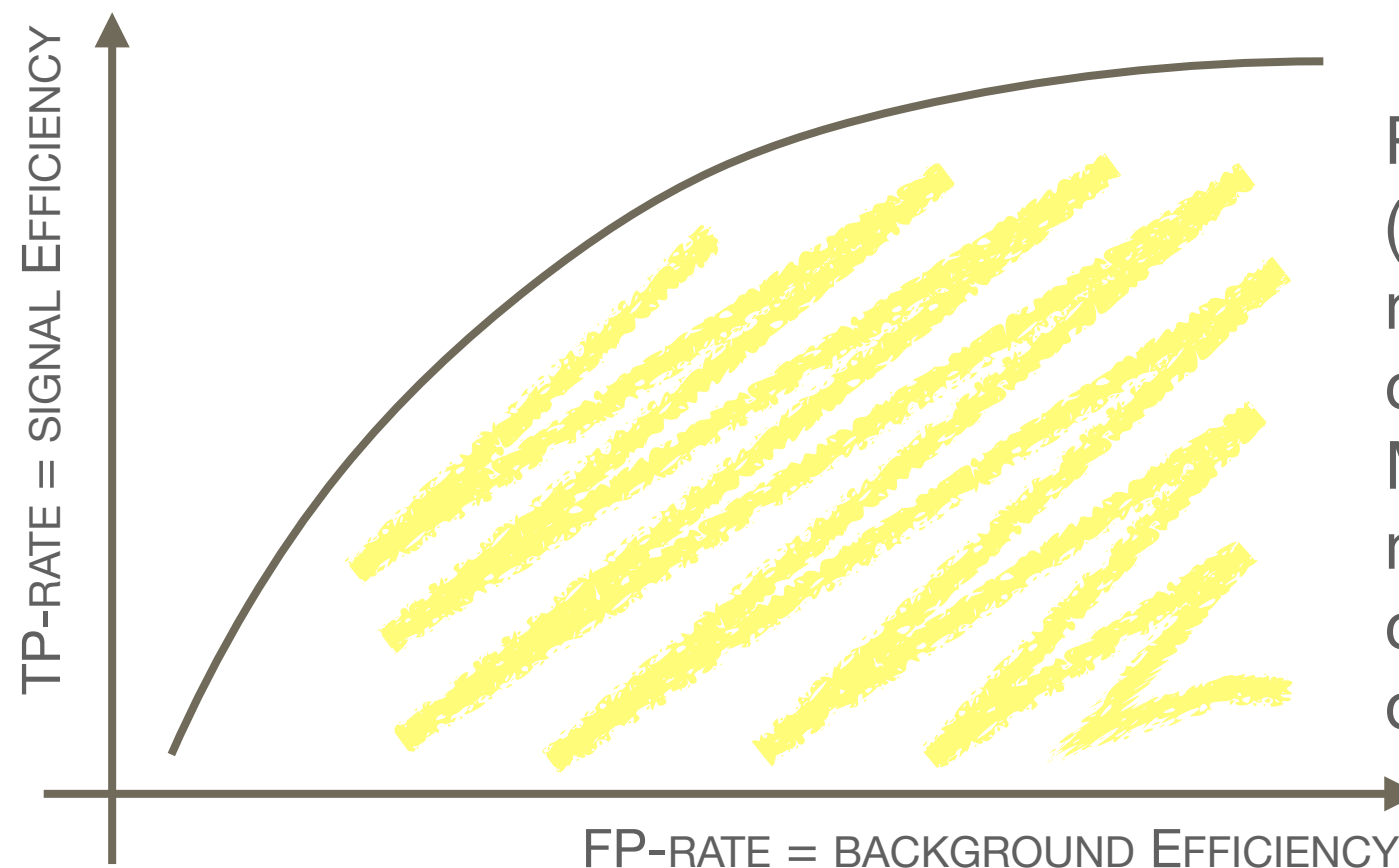
Specificity = $\frac{TN}{TN + FP}$

FP rate = 1 - specificity = $1 - \frac{TN}{TN + FP} = \frac{FP}{TN + FP}$

FP rate = background efficiency

TP rate = sensitivity = $\frac{TP}{TP + FN}$

TP rate = signal efficiency



ROC-AUC

(Area Under Curve)
misura l'area sotto la curva ROC;
Maggiore è l'AUC, migliore è la capacità di classificazione corretta dell'algoritmo

THE ROC (RECEIVER OPERATING CHARACTERISTIC): TPRATE VS FPRATE

Signal events = $TP + FN$ # Background events = $FP + TN$

Signal efficiency: $\frac{TP}{TP + FN}$ Sensitivity = $\frac{TP}{TP + FN}$

Background efficiency: $\frac{FP}{FP + TN}$

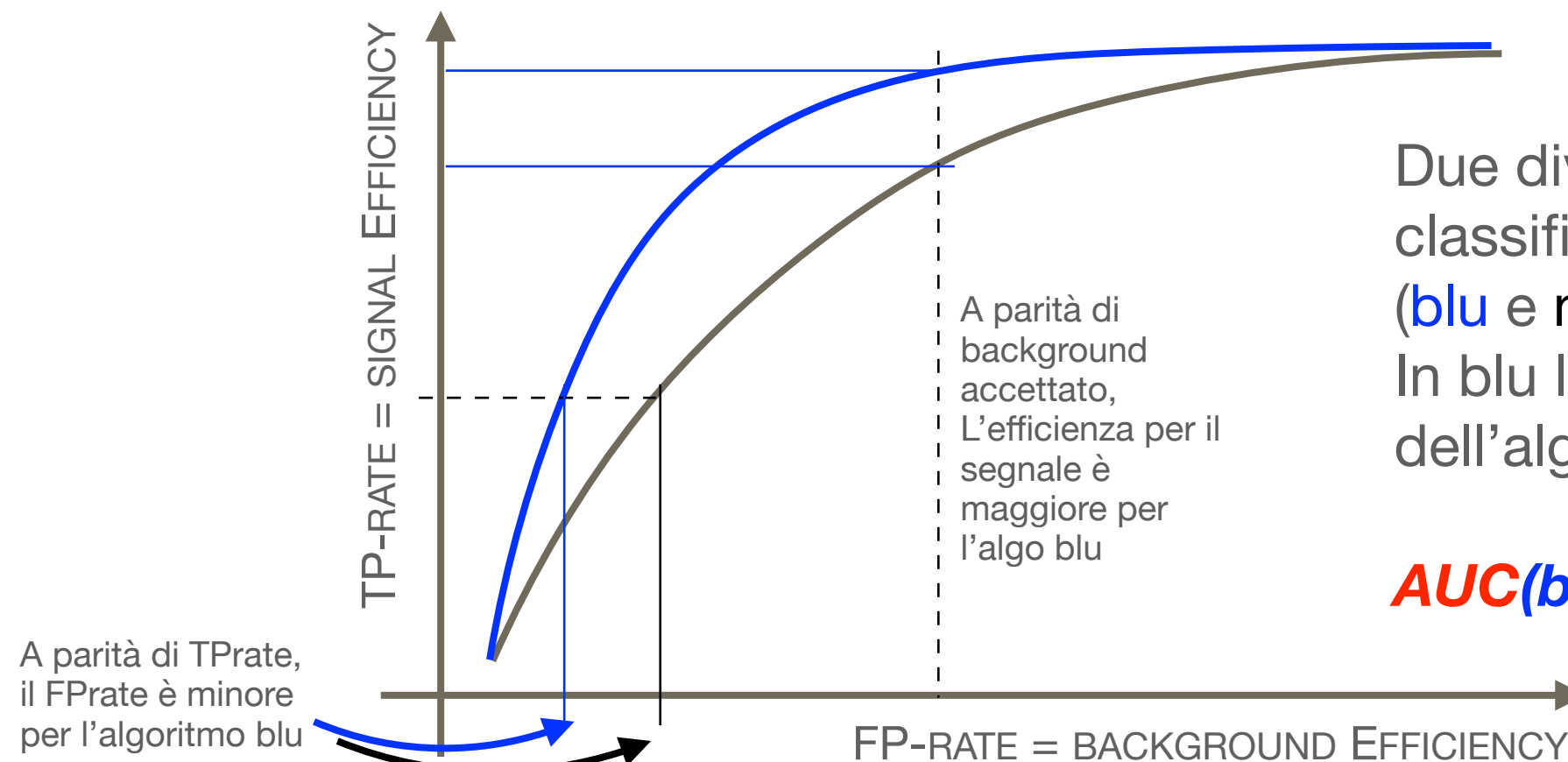
Specificity = $\frac{TN}{TN + FP}$

FPrate = 1 - specificity = $1 - \frac{TN}{TN + FP} = \frac{FP}{TN + FP}$

FPrate = background efficiency

TPrate = sensitivity = $\frac{TP}{TP + FN}$

TPrate = signal efficiency



Due diverso algoritmi di classificazione (blu e nero) a confronto. In blu la ROC dell'algoritmo migliore

$AUC(blue) > AUC(nero)$

THE ROC (RECEIVER OPERATING CHARACTERISTIC): TPrate VS FPrate

Signal events = $TP + FN$ # Background events = $FP + TN$

Signal efficiency: $\frac{TP}{TP + FN}$ Sensitivity = $\frac{TP}{TP + FN}$

Background efficiency: $\frac{FP}{FP + TN}$

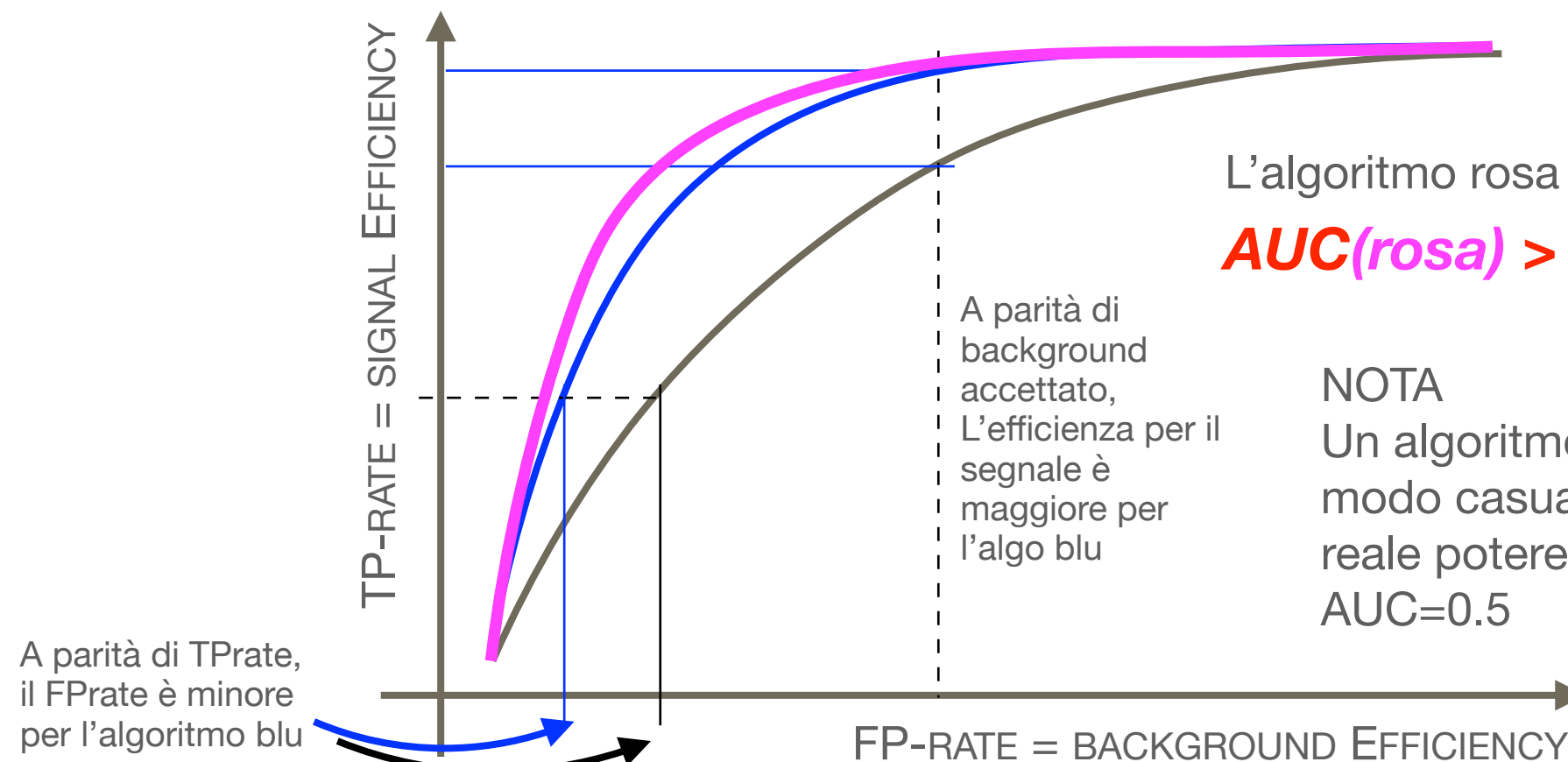
Specificity = $\frac{TN}{TN + FP}$

FPrate = 1 - specificity = $1 - \frac{TN}{TN + FP} = \frac{FP}{TN + FP}$

FPrate = background efficiency

TPrate = sensitivity = $\frac{TP}{TP + FN}$

TPrate = signal efficiency



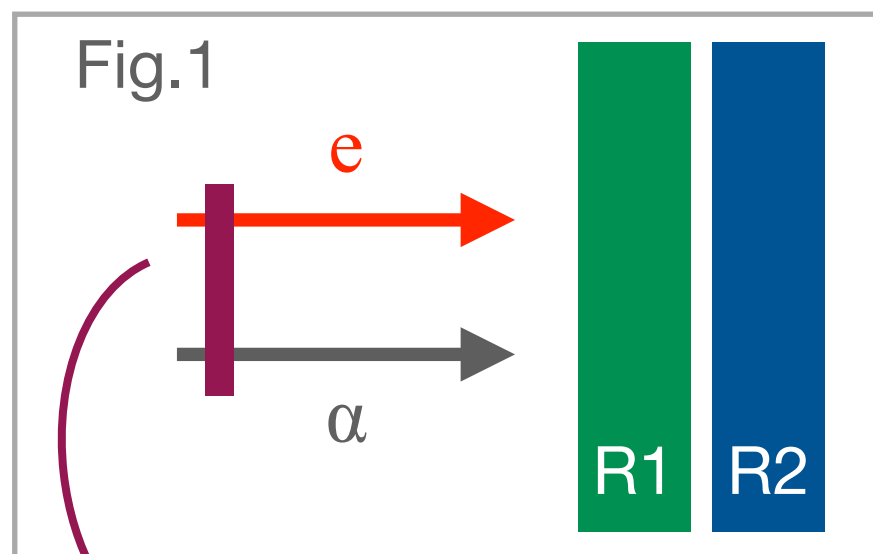
L'algoritmo rosa ha prestazioni ancora migliori
 $AUC(\text{rosa}) > AUC(\text{blu}) > AUC(\text{nero})$

NOTA

Un algoritmo che classifica in modo casuale, cioè non ha nessun reale potere discriminante, ha $AUC=0.5$

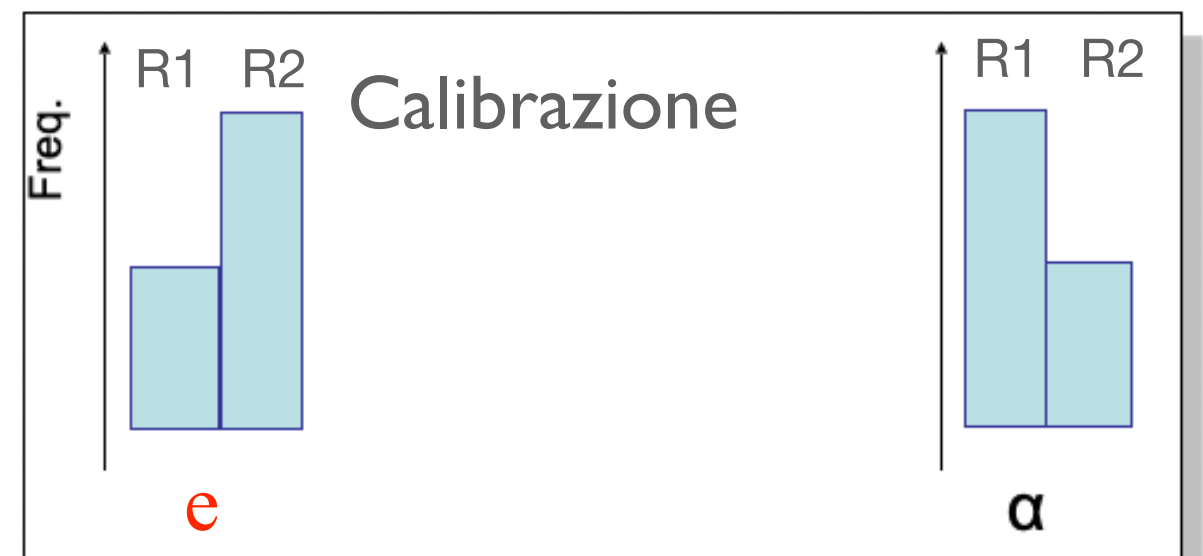
CALCOLO COMBINATORIO E PROBABILITÀ

- Un altro esempio: consideriamo un apparato sperimentale costituito da due rivelatori consecutivi che rivelano l'interazione di particelle (fig.1). Sull'apparato sperimentale arrivano 2 tipi di particelle: elettroni e particelle alpha (α). Se i due rivelatori hanno una diversa "risposta" (sensibilità) alle due tipologie di particelle posso determinare la composizione del fascio (quanti e, quante α) ?



Tagger: rivelatore ausiliario che indica il passaggio di una particella del fascio con la stessa probabilità per elettroni e alpha

Precedentemente avrò calibrato il mio rivelatore su fasci di particelle di soli elettroni e sole α per determinarne la risposta di R1 e R2 ai diversi tipi di particella (fig.2)



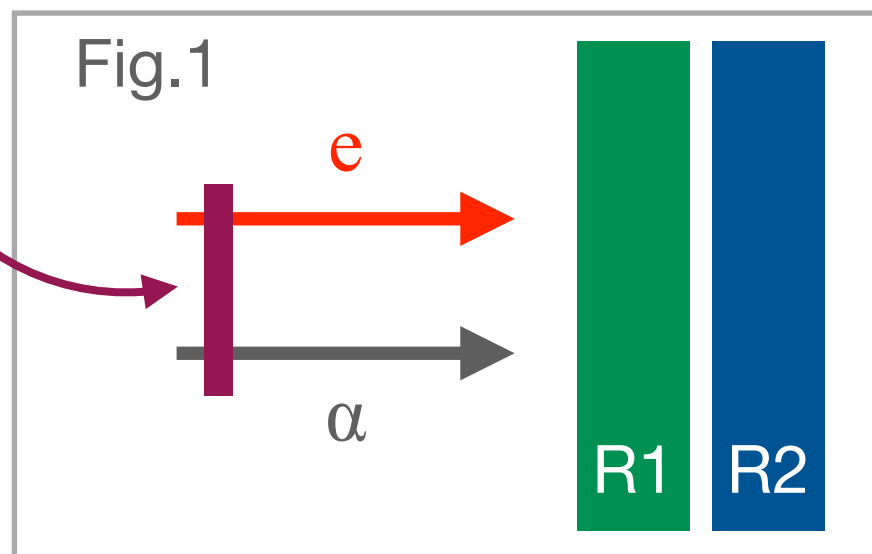
Immaginiamo di avere
 $N=10000$ particelle segnalate dal tagger
 $\Rightarrow$ osserverò i conteggi in ciascun rivelatore,
indipendentemente oppure in correlazione

Cause, classi di eventi: e, α
Effetti: segnale in R1, segnale in R2
 $P(R1/e) = 0.1$ $P(R1/\alpha) = 0.9$
 $P(R2/e) = 0.8$ $P(R2/\alpha) = 0.05$

CALCOLO COMBINATORIO E PROBABILITÀ

- R1 ha efficienza 10 % per elettroni, 90% per particelle alpha
- R2 ha efficienza 80 % per elettroni, 5% per particelle alpha
- 4 esiti possibili:
 - R1 && !R2
 - R1 && R2
 - !R1 && !R2
 - !R1 && R2

Tagger: indica il passaggio di una particella del fascio con la stessa probabilità per elettroni e alpha



Cause, classi di eventi: e, α

Effetti: segnale in R1, segnale in R2

$$P(R1/e) = 0.1 \quad P(R1/\alpha) = 0.9$$

$$P(R2/e) = 0.8 \quad P(R2/\alpha) = 0.05$$

$$f(e) = 0.4, f(\alpha) = 0.6 \text{ note}$$

Per N=10000 particelle segnalate dal tagger osservo

Risultato di una singola simulazione MC (1 solo esp.)

$$N(R1 \ \&\& \ !R2) = 6875$$

$$N(R1 \ \&\& \ R2) = 559$$

$$N(!R1 \ \&\& \ !R2) = 1059$$

$$N(!R1 \ \&\& \ R2) = 10000 - 6875 - 559 - 1059 = 1507$$

CALCOLO COMBINATORIO E PROBABILITÀ

- R1 ha efficienza 10 % per elettroni, 90% per particelle alpha
- R2 ha efficienza 80 % per elettroni, 5% per particelle alpha
- 4 esiti possibili:

Cause, classi di eventi: e, α

Effetti: segnale in R1, segnale in R2

$$P(R1|e) = 0.1 \quad P(R1|\alpha) = 0.9$$

$$P(R2|e) = 0.8 \quad P(R2|\alpha) = 0.05$$

- R1 && !R2
- R1 && R2
- !R1 && !R2
- !R1 && R2

Per N=10000 particelle segnalate dal tagger osservo

Risultato di una singola simulazione MC (1 solo esp.)

$$N(R1 \&\& !R2) = 6875$$

$$N(R1 \&\& R2) = 559$$

$$N(!R1 \&\& !R2) = 1059$$

$$N(!R1 \&\& R2) = 10000 - 6875 - 559 - 1059 = 1507$$

- Se consideriamo solo $N(R1) = 7434$ e $N(R2) = 2066$

$$N(R1) = N \times P(R1) = N \times (f_e \times P(R1|e)) + N \times (f_\alpha \times P(R1|\alpha))$$

Teo prob totale

$$N(R2) = N \times P(R2) = N \times (f_e \times P(R2|e)) + N \times (f_\alpha \times P(R2|\alpha))$$

Teo prob totale

- Due eq. in due incognite, f_e e f_α , se il processo fosse deterministico

- In realta' dal momento che $f_\alpha = 1 - f_e$ da ogni eq. si ricava f_e e di conseguenza f_α

$$P(e | R2) = P(R2|e)P(e)/P(R2)$$

$$P(\alpha | R1) = P(R1|\alpha)P(\alpha)/P(R1)$$

CALCOLO COMBINATORIO E PROBABILITÀ

- R1 ha efficienza 10 % per elettroni, 90% per particelle alpha
- R2 ha efficienza 80 % per elettroni, 5% per particelle alpha
- 4 esiti possibili:

Cause, classi di eventi: e, α

Effetti: segnale in R1, segnale in R2

$$P(R1|e) = 0.1 \quad P(R1|\alpha) = 0.9$$

$$P(R2|e) = 0.8 \quad P(R2|\alpha) = 0.05$$

- R1 && !R2
- R1 && R2
- !R1 && !R2
- !R1 && R2

Per N=10000 particelle segnalate dal tagger osservo

Risultato di una singola simulazione MC (1 solo esp.)

$$N(R1 \&\& !R2) = 6875$$

$$N(R1 \&\& R2) = 559$$

$$N(!R1 \&\& !R2) = 1059$$

$$N(!R1 \&\& R2) = 10000 - 6875 - 559 - 1059 = 1507$$

- $P(R1 \&\& !R2 | e) = 0.1 \times (1 - 0.8) = 0.02$
- $P(R1 \&\& R2 | e) = 0.1 \times 0.8 = 0.08$
- $P(!R1 \&\& !R2 | e) = (1 - 0.1) \times (1 - 0.8) = 0.18$
- $P(!R1 \&\& R2 | e) = (1 - 0.1) \times 0.8 = 0.72$

Somma = 100% Teo prob totale

- $P(R1 \&\& !R2 | \alpha) = 0.9 \times (1 - 0.05) = 0.855$
- $P(R1 \&\& R2 | \alpha) = 0.9 \times 0.05 = 0.045$
- $P(!R1 \&\& !R2 | \alpha) = (1 - 0.9) \times (1 - 0.05) = 0.095$
- $P(!R1 \&\& R2 | \alpha) = (1 - 0.9) \times 0.05 = 0.005$

Somma = 100%

$$p(e | R1\&\&!R2) = 0.02 \times P(e) / p(R1\&\&!R2) = 0.02 P(e) / (0.02 P(e) + 0.855 P(\alpha))$$

$$p(e | R1\&\&R2) = 0.08 \times P(e) / p(R1\&\&R2) = 0.08 P(e) / (0.08 P(e) + 0.045 P(\alpha))$$

$$p(e | !R1\&\&!R2) = 0.18 \times P(e) / p(!R1\&\&!R2) = 0.18 P(e) / (0.18 P(e) + 0.095 P(\alpha))$$

$$p(e | !R1\&\&R2) = 0.72 \times P(e) / p(!R1\&\&R2) = 0.72 P(e) / (0.72 P(e) + 0.005 P(\alpha))$$

CALCOLO COMBINATORIO E PROBABILITÀ

OSSERVAZIONE

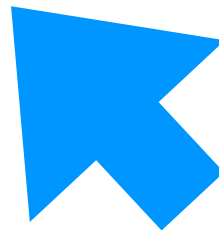
$p(R1 \& \bar{R2})$ e' stimato da $6875/10000 = 0.68$
 $p(R1 \& R2)$ e' stimato da $559/100000 = 0.00559$
 $p(\bar{R1} \& \bar{R2})$ e' stimato da $1059/100000 = 0.01059$
 $p(\bar{R1} \& R2)$ e' stimato da $1507/100000 = 0.01507$

Cause, classi di eventi: e, a

Effetti: segnale in R1, segnale in R2

$$P(R1/e) = 0.1 \quad P(R1/a) = 0.9$$

$$P(R2/e) = 0.8 \quad P(R2/a) = 0.05$$



Per $N=10000$ particelle segnalate dal tagger osservo

Risultato di una singola simulazione MC (1 solo esp.)

$$N(R1 \& \bar{R2}) = 6875$$

$$N(R1 \& R2) = 559.$$

$$N(\bar{R1} \& \bar{R2}) = 1059$$

$$N(\bar{R1} \& R2) = 1507$$

$$--->>> N(R1) = 7434$$

$$--->>> N(R2) = 2066$$

$$p(e | R1 \& \bar{R2}) = 0.02 \times P(e) / p(R1 \& \bar{R2}) = 0.02 P(e) / (0.02 P(e) + 0.855 P(a))$$

$$p(e | R1 \& R2) = 0.08 \times P(e) / p(R1 \& R2) = 0.08 P(e) / (0.08 P(e) + 0.045 P(a))$$

$$p(e | \bar{R1} \& \bar{R2}) = 0.18 \times P(e) / p(\bar{R1} \& \bar{R2}) = 0.18 P(e) / (0.18 P(e) + 0.095 P(a))$$

$$p(e | \bar{R1} \& R2) = 0.72 \times P(e) / p(\bar{R1} \& R2) = 0.72 P(e) / (0.72 P(e) + 0.005 P(a))$$

SIMULAZIONE

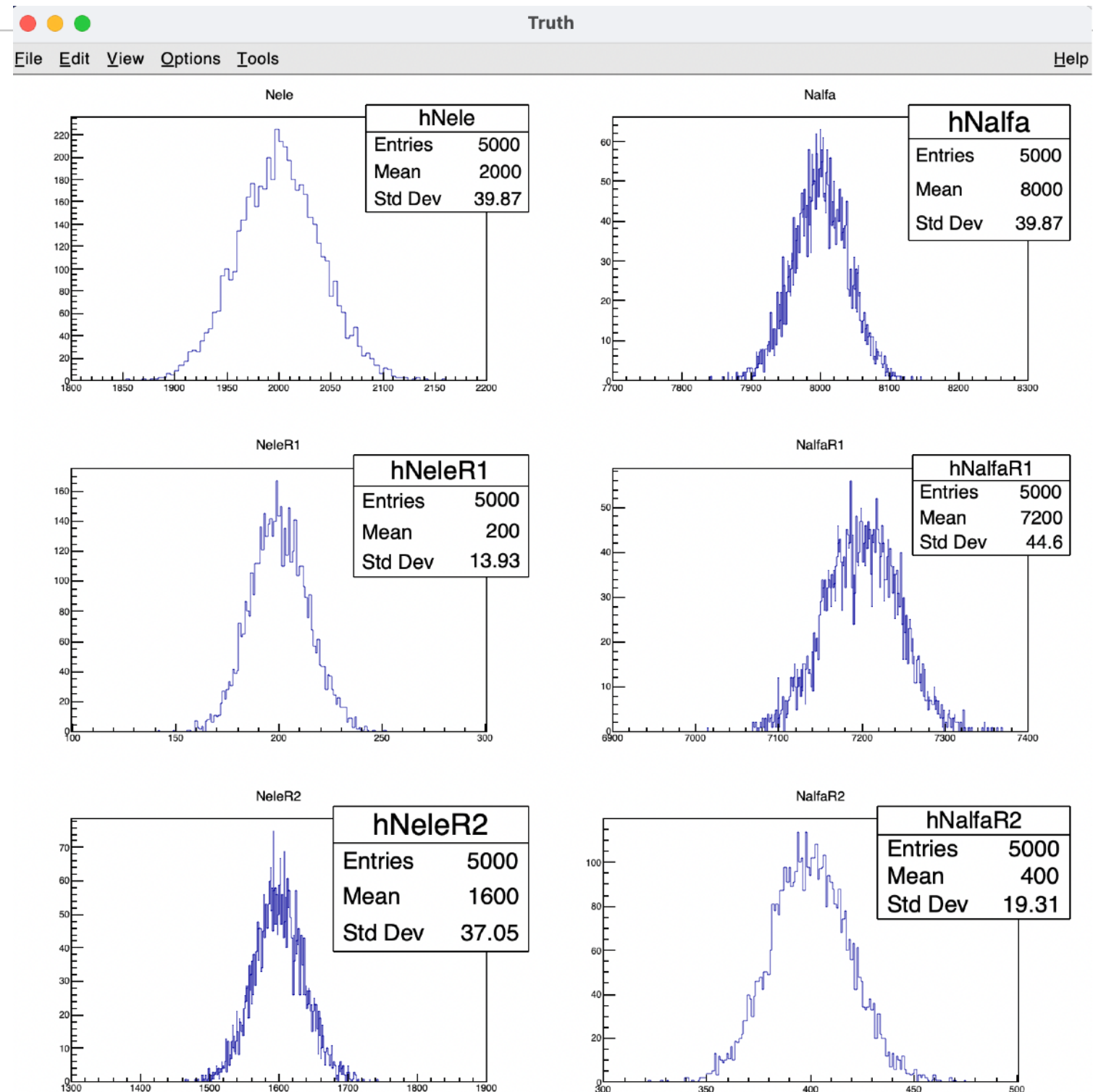
- 5000 esperimenti
 - In ciascuno ho 10000 particelle
 - Nella mia simulazione la frazione di e- ne; fascio e' 20%

La verita' MC:

Quanti elettroni (alfa) sono prodotti in ogni esperimento ?

Quanti elettroni (alfa) danno segnale in R1 ?

Quanti elettroni (alfa) danno segnale in R2 ?



SIMULAZIONE

- 5000 esperimenti
 - In ciascuno ho 10000 particelle

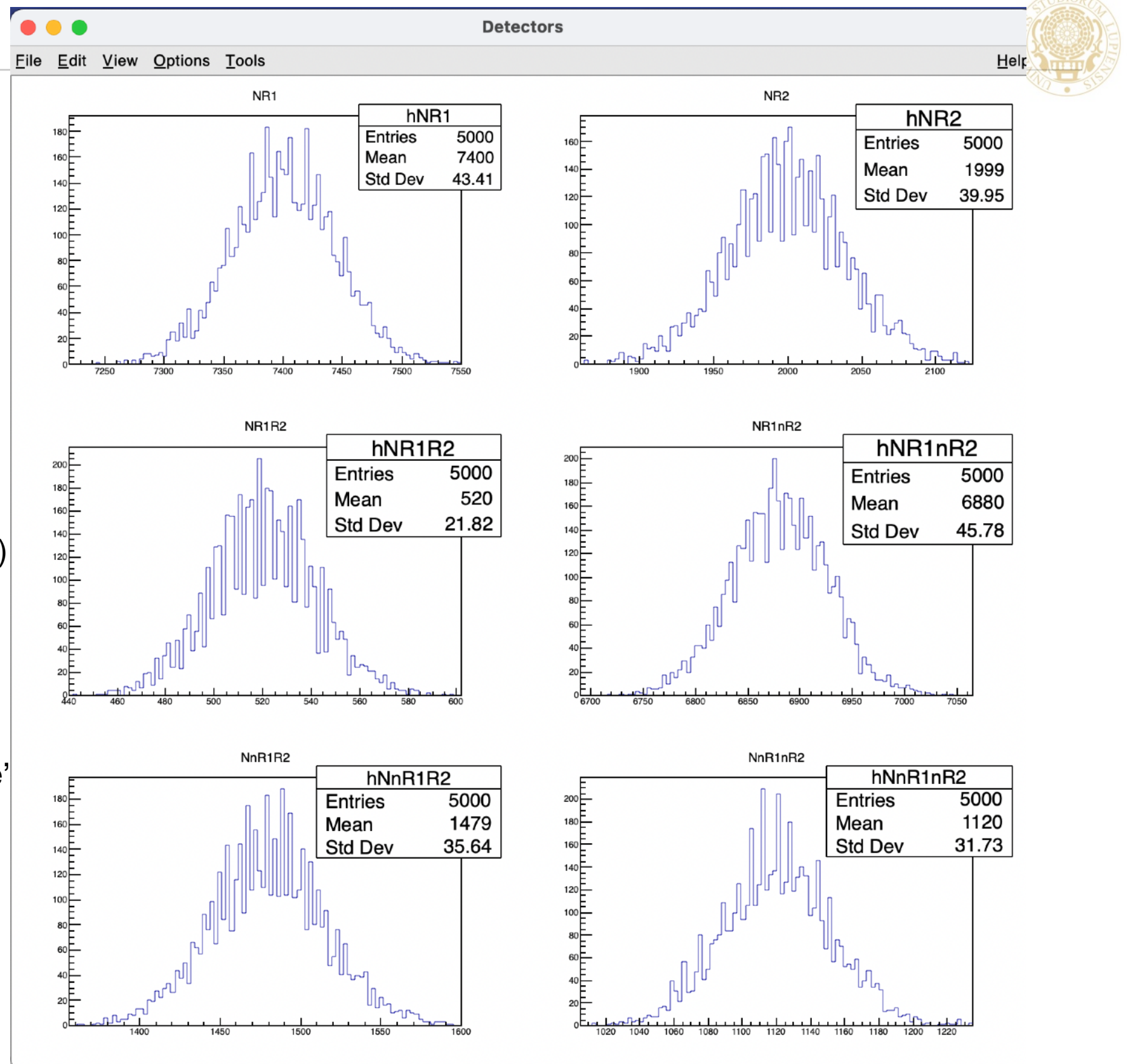
Le misure dei rivelatori:

Quanti segnali in tutto ci sono in R1 (R2) in ogni esperimento ?

Quante particelle danno segnali sia in R1 che in R2 ?

Quante particelle danno segnale solo in R1 (R2) ?

Quante particelle non danno segnale ne' R1 ne' in R2 ?



SIMULAZIONE

- 5000 esperimenti
 - In ciascuno ho 10000 particelle

xBayes: $P(e|nR1) = 0.692308$ $P(a|nR1) 0.307692$

xBayes: $P(e| R1) = 0.027027$ $P(a| R1) 0.972973$

xBayes: $P(e|nR1 R2) = 0.972973$ $P(a|nR1 R2) 0.027027$

xBayes: $P(e| R1 R2) = 0.307692$ $P(a| R1 R2) 0.692308$

xBayes: $P(e|nR1nR2) = 0.321429$ $P(a|nR1 R2) 0.678571$

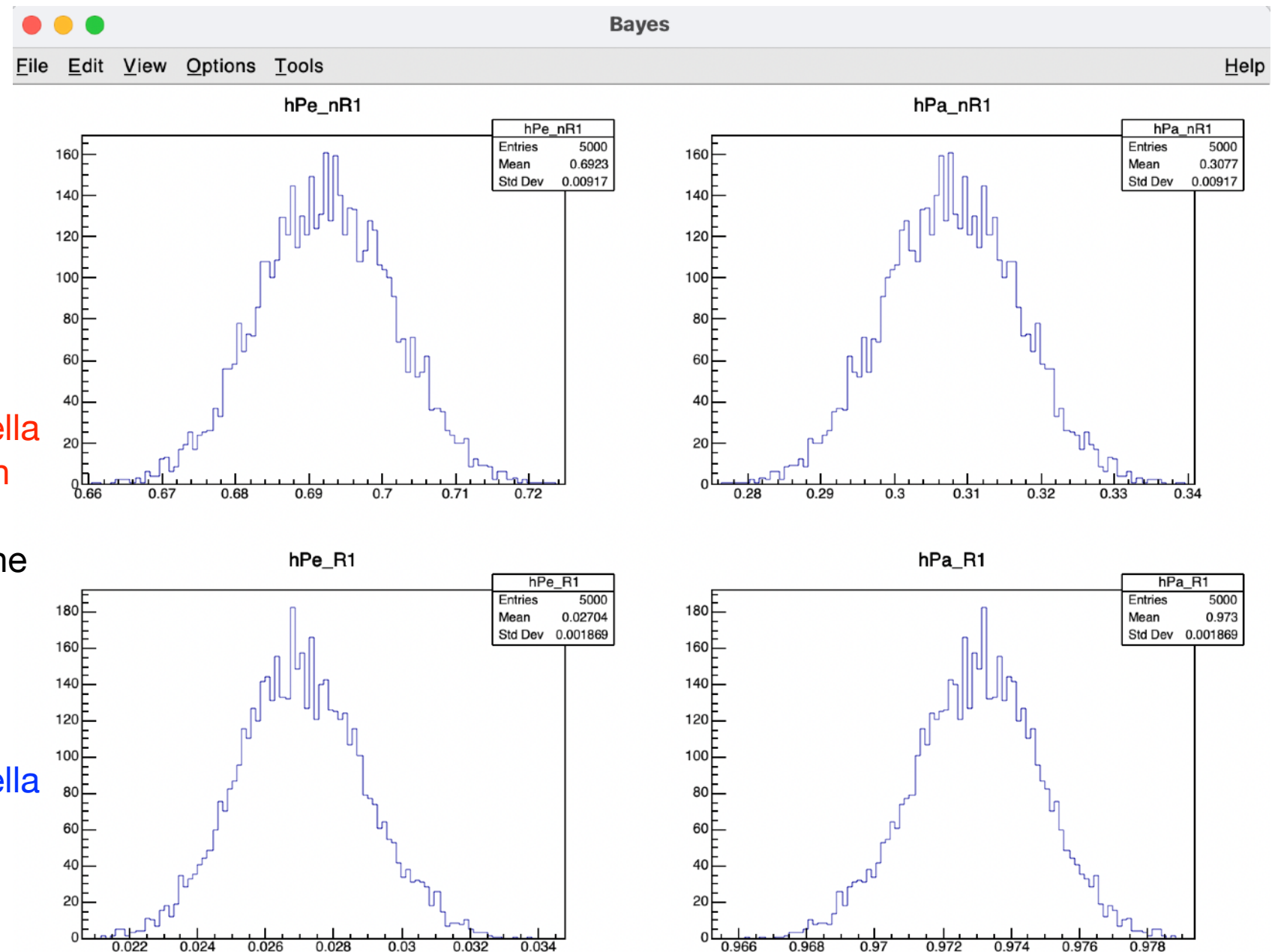
xBayes: $P(e| R1nR2) = 0.00581395$ $P(a| R1nR2) 0.994186$

Predizioni del Teorema di Bayes a confronto con la simulazione:

Qual e' la probabilita' che una particella che non produce segnale in R1 sia un elettrone (alfa) ?

(MC: quale frazione delle particelle che non producono segnale in R1 corrisponde a elettroni (alfa) [uso al verita' MC]).

Qual e' la probabilita' che una particella che produce segnale in R1 sia un elettrone (alfa) ?



SIMULAZIONE

xBayes: $P(e|nR1) = 0.692308$ $P(a|nR1) = 0.307692$

xBayes: $P(e| R1) = 0.027027$ $P(a| R1) = 0.972973$

xBayes: $P(e|nR1 R2) = 0.972973$ $P(a|nR1 R2) = 0.027027$

xBayes: $P(e| R1 R2) = 0.307692$ $P(a| R1 R2) = 0.692308$

xBayes: $P(e|nR1nR2) = 0.321429$ $P(a|nR1 R2) = 0.678571$

xBayes: $P(e| R1nR2) = 0.00581395$ $P(a| R1nR2) = 0.994186$

- 5000 esperimenti
- In ciascuno ho 10000 particelle

Predizioni del Teorema di Bayes a confronto con la simulazione:

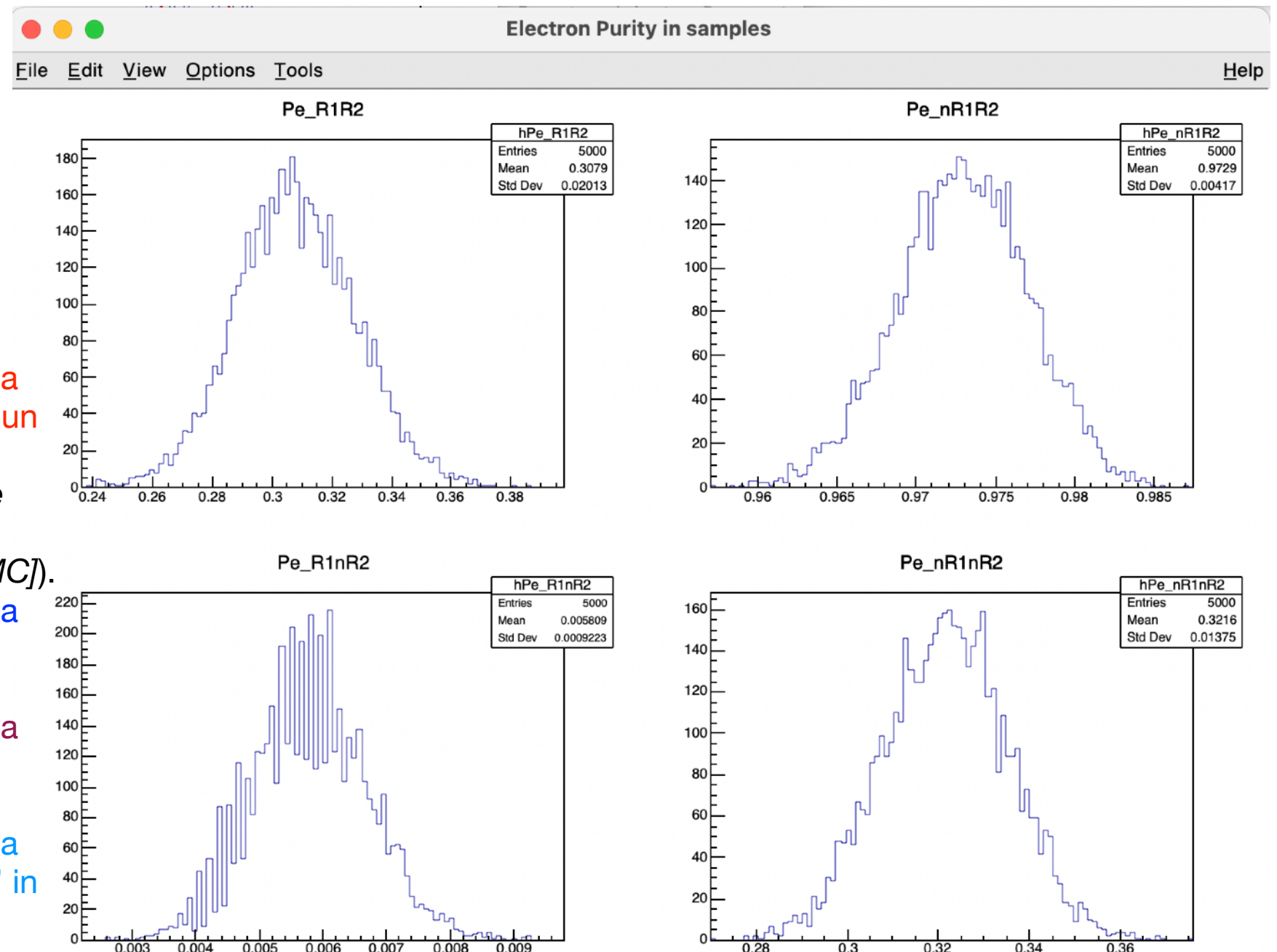
Qual e' la probabilita' che una particella che produce segnale in R1 e in R2 sia un elettrone ?

(MC: quale frazione delle particelle che producono segnali in R1 e in R2 corrisponde a elettroni [uso al verita' MC]).

Qual e' la probabilita' che una particella che produce segnale solo in R2 sia un elettrone ?

Qual e' la probabilita' che una particella che produce segnale solo in R1 sia un elettrone ?

Qual e' la probabilita' che una particella che non produce segnale ne' in R1 ne' in R2 sia un elettrone ?



SIMULAZIONE

xBayes: $P(e|nR1) = 0.692308$ $P(a|nR1) 0.307692$

xBayes: $P(e| R1) = 0.027027$ $P(a| R1) 0.972973$

xBayes: $P(e|nR1 R2) = 0.972973$ $P(a|nR1 R2) 0.027027$

xBayes: $P(e| R1 R2) = 0.307692$ $P(a| R1 R2) 0.692308$

xBayes: $P(e|nR1nR2) = 0.321429$ $P(a|nR1 R2) 0.678571$

xBayes: $P(e| R1nR2) = 0.00581395$ $P(a| R1nR2) 0.994186$

- 5000 esperimenti
 - In ciascuno ho 10000 particelle

Predizioni del Teorema di Bayes a confronto con la simulazione:

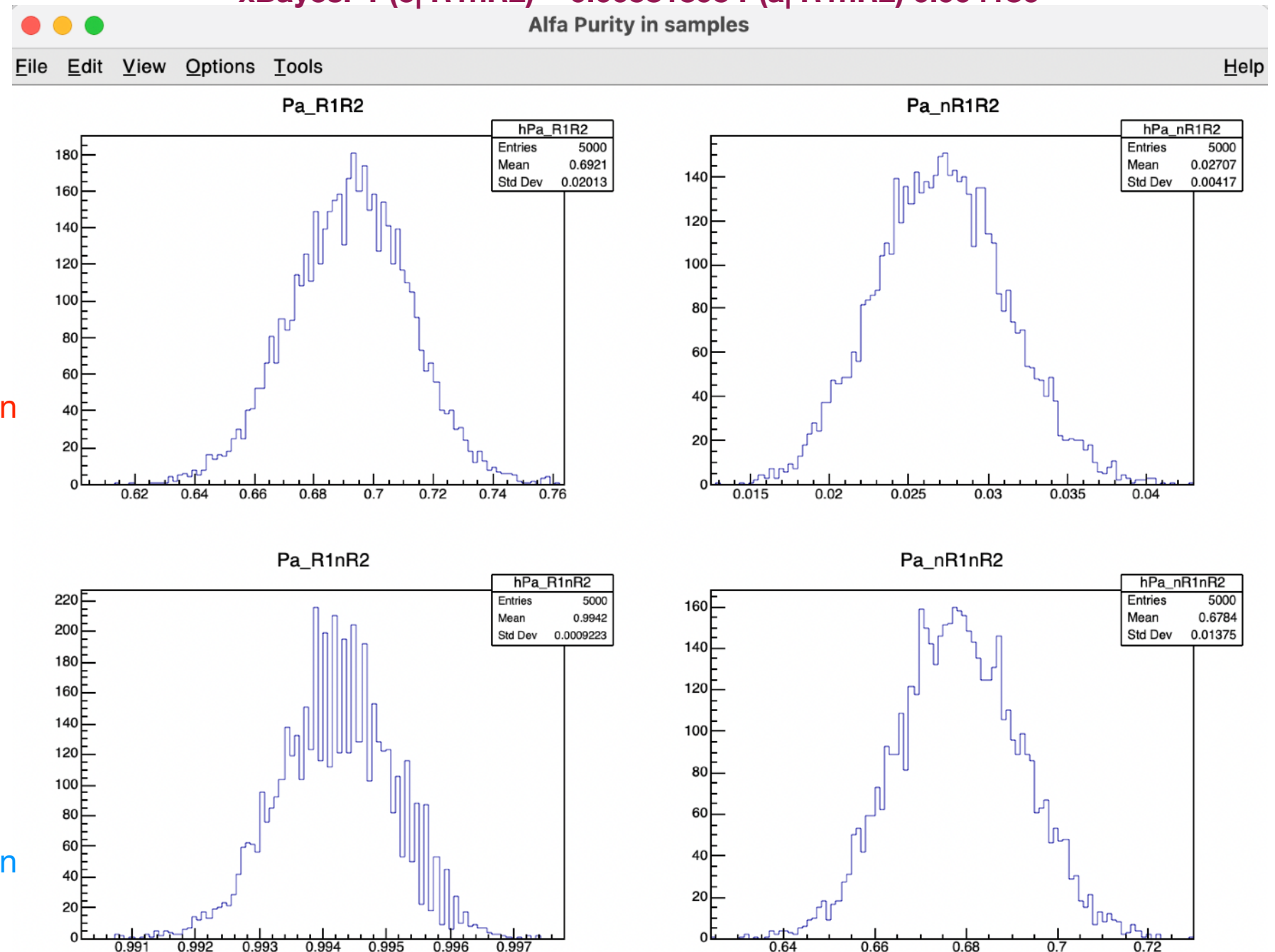
Qual e' la probabilita' che una particella che produce segnale in R1 e in R2 sia un alfa ?

(MC: quale frazione delle particelle che producono segnali in R1 e in R2 corrisponde ad alfa [uso al verita' MC]).

Qual e' la probabilita' che una particella che produce segnale solo in R2 sia una alfa ?

Qual e' la probabilita' che una particella che produce segnale solo in R1 sia un alfa ?

Qual e' la probabilita' che una particella che non produce segnale ne' in R1 ne' in R2 sia un alfa ?

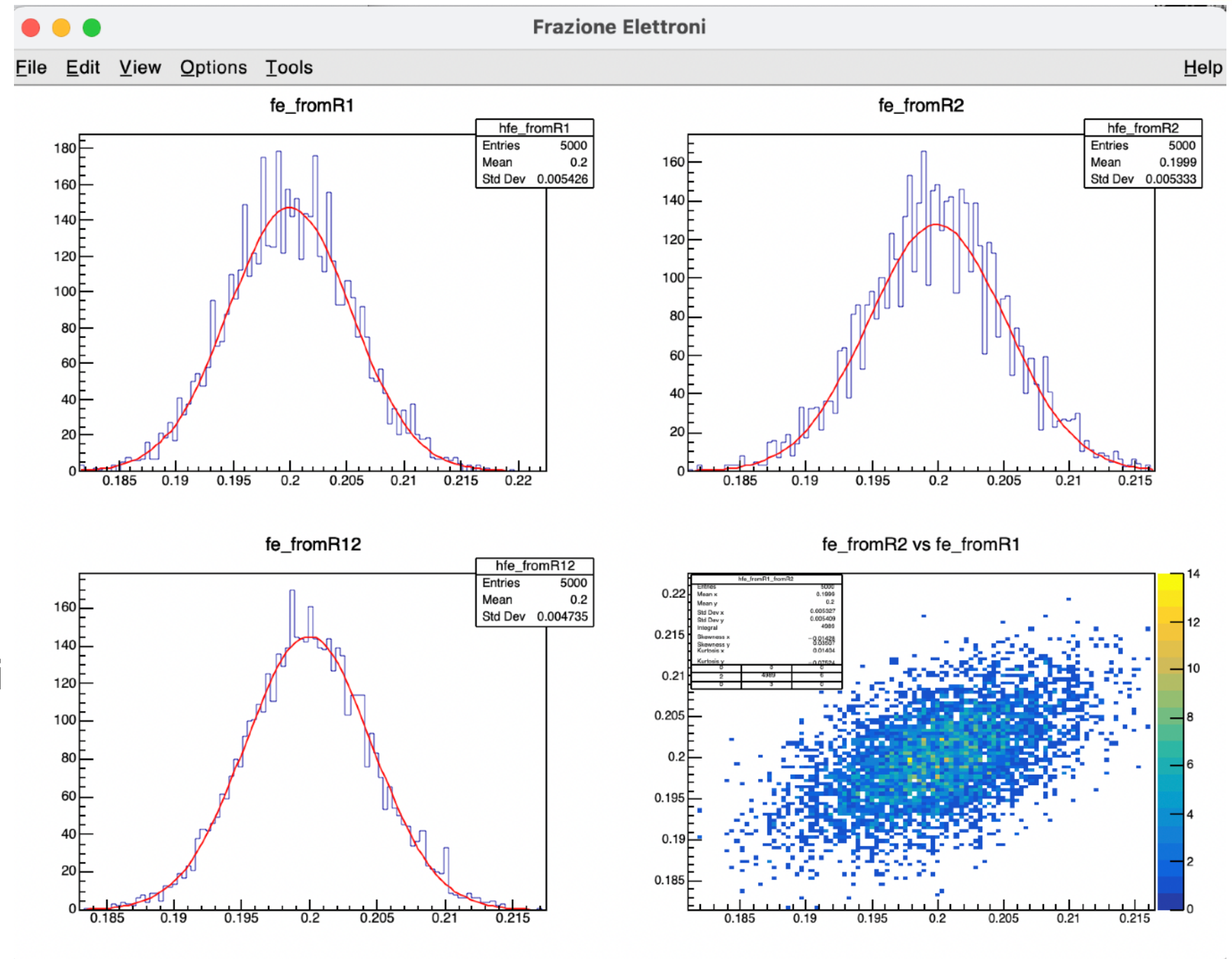


SIMULAZIONE

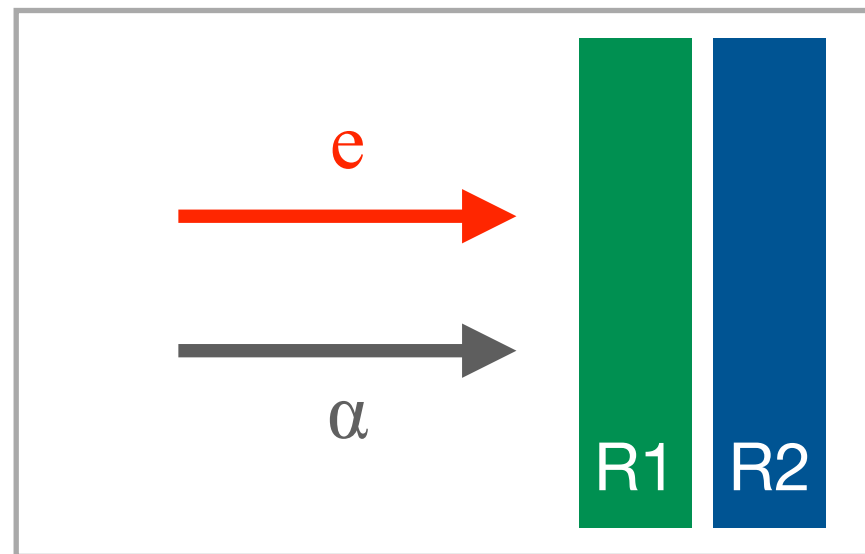
- 5000 esperimenti
 - In ciascuno ho 10000 particelle

Osservando solo il numero di segnali in R1 (R1), note le efficienze di R1 (R2) sia per elettroni che per alfa posso stimare in ogni esperimento la frazione di elettroni nel fascio.

Sono informazioni indipendenti ?



CALCOLO COMBINATORIO E PROBABILITÀ



Il mio esperimento: decadimento ignoto
Conto (in un modo indipendente da R1 e R2) i decadimenti totali
Conto le risposte di R1
Mi chiedo quanti decadimenti abbiano prodotto particelle α ?
Voglio insomma determinare $P(\alpha | R1)$

- Le conclusioni del nostro esperimento dipendono dalla conoscenza a priori del processo in studio?
- Con che probabilità le particelle che interagiranno sul rivelatore 1 sono alpha?

CALCOLO COMBINATORIO E PROBABILITÀ

$$P(B_k | A_i) = \frac{P(A_i | B_k) P(B_k)}{\sum_j P(A_i | B_j) P(B_j)}$$

- NOTA: per poter stimare la probabilità che la misura (effetto) sia attribuibile ad una delle possibili cause devo introdurre una stima della probabilità che una delle cause sia presente ($P(B_k)$).
- Nell'approccio Bayesiano alla misura ho sempre una probabilità a priori ($P(B_k)$) e una probabilità a posteriori ($P(B_k|A)$).
- La misura altera la conoscenza del fenomeno in studio cambiando la convinzione su quali sono le cause che sono alla base dell'universo che osserviamo.
- La assoluta ignoranza di un processo oggetto di studio comporta la necessità di assegnare a tutte le possibili cause la stessa probabilità.

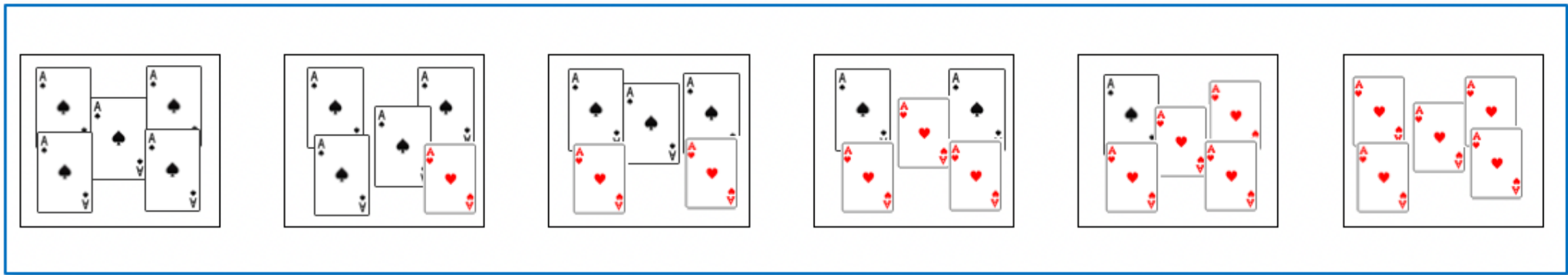
CALCOLO COMBINATORIO E PROBABILITÀ

$$P(B_k | A_i) = \frac{P(A_i | B_k) P(B_k)}{\sum_j P(A_i | B_j) P(B_j)}$$

- La assoluta ignoranza di un processo oggetto di studio comporta la necessità di assegnare a tutte le possibili cause la stessa probabilità.
- Consideriamo quindi tutte le possibili cause sono equiprobabili
 - $P(B_k)=1/N$; $k=1,\dots,N$
- Tutte le nostre conoscenze sono frutto di probabilità a posteriori.
- Lo stesso concetto di probabilità assoluta non esiste, è solo un assioma matematico. In termini pratici possiamo parlare solo di probabilità a posteriori.

CALCOLO COMBINATORIO E PROBABILITÀ

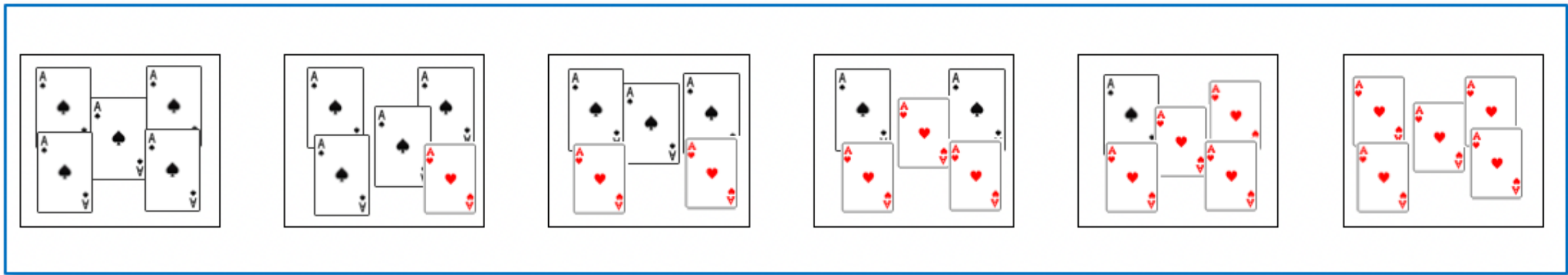
Il problema delle sei buste



- Quant'è la probabilità di estrarre una carta rossa da una delle buste prese a caso?
- Supponiamo di aver estratto una carta rossa da una busta e di averla rimessa dentro la stessa busta, quant'è, ora, la probabilità di estrarre una carta rossa dalla stessa busta?

CALCOLO COMBINATORIO E PROBABILITÀ

Il problema delle sei buste



- Quant'è la probabilità di estrarre una carta rossa da una delle buste prese a caso?

6 cause, ossia sei buste, equiprobabili;

$$p(\text{rossa}, B1) = 0 \times 1/6$$

$$p(\text{rossa}, B2) = 1/5 \times 1/6$$

$$p(\text{rossa}, B3) = 2/5 \times 1/6$$

$$p(\text{rossa}, B4) = 3/5 \times 1/6$$

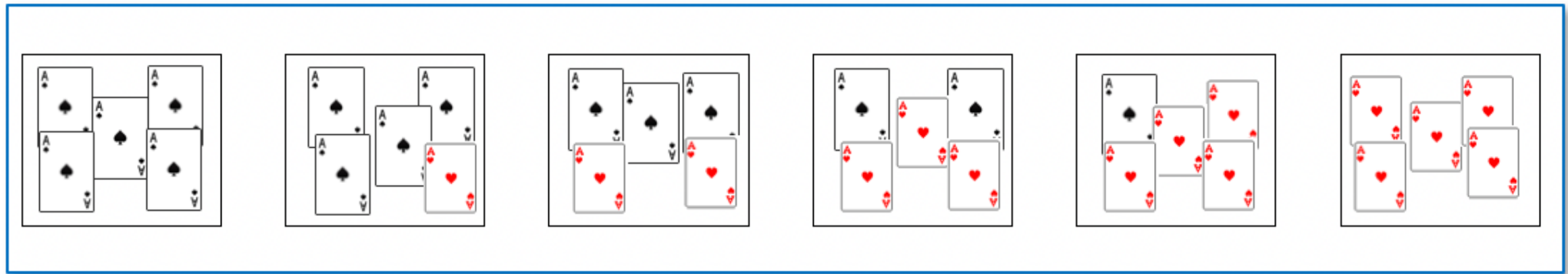
$$p(\text{rossa}, B5) = 4/5 \times 1/6$$

$$p(\text{rossa}, B6) = 5/5 \times 1/6$$

$$p(\text{rossa}, \text{any } B) = 15/5 \times 1/6 = 1/2$$

CALCOLO COMBINATORIO E PROBABILITÀ

Il problema delle sei buste



- Supponiamo di aver estratto una carta rossa da una busta e di averla rimessa dentro la stessa busta, quant'è, ora, la probabilità di estrarre una carta rossa dalla stessa busta?

$$p(\text{rossa}, B2) = 1/5 \times 1/5$$

$$p(\text{rossa}, B3) = 2/5 \times 1/5$$

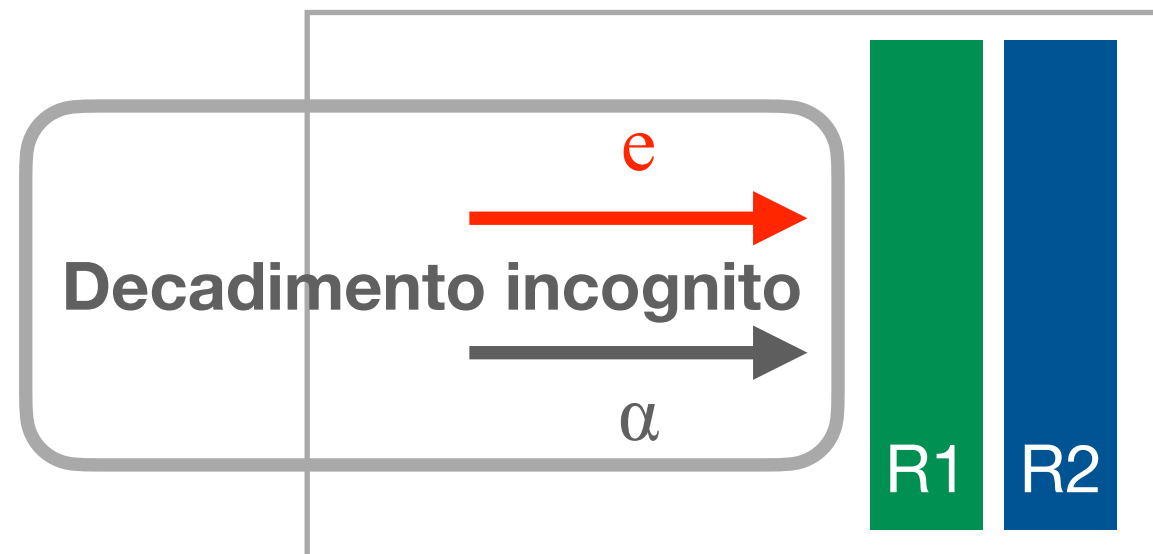
$$p(\text{rossa}, B4) = 3/5 \times 1/5$$

$$p(\text{rossa}, B5) = 4/5 \times 1/5$$

$$p(\text{rossa}, B6) = 5/5 \times 1/5$$

$$p(\text{rossa}, \text{any } B \text{ but } B1) = 15/5 \times 1/5 = 3/5$$

CALCOLO COMBINATORIO E PROBABILITÀ



Le conclusioni del nostro esperimento dipendono dalla conoscenza a priori del processo in studio?

Con che probabilità le particelle che interagiranno sul rivelatore 1 sono alpha?

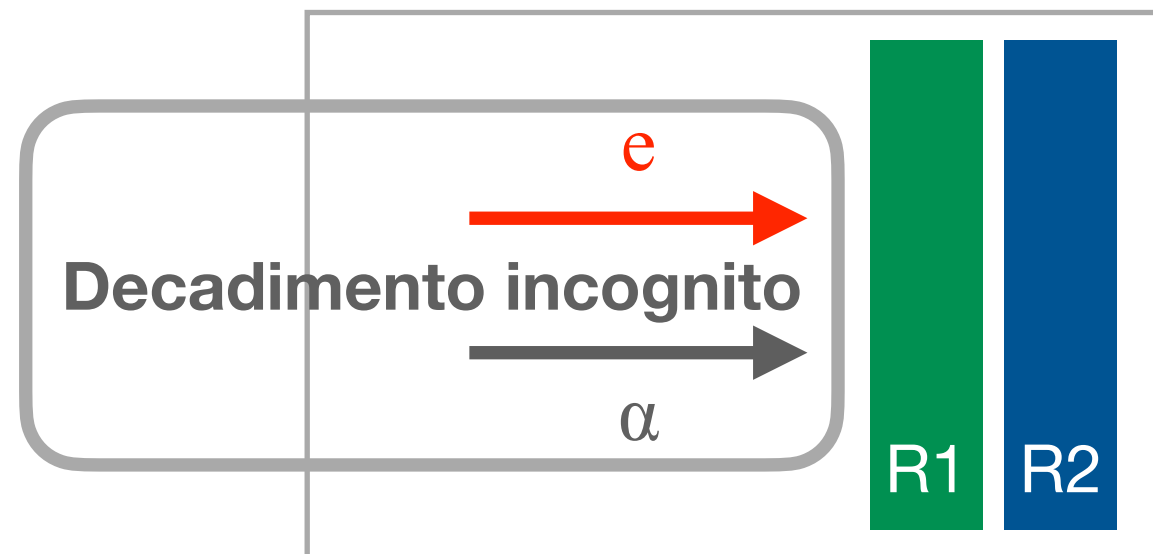
$$P(\alpha | R_1) = P(R_1 | \alpha) P(\alpha) / P(R_1)$$

Probabilità a priori uniforme = 0.5 per ogni tipo di particella

$$P(R_1) = P(R_1 | \alpha) P(\alpha) + P(R_1 | e) P(e) = 0.9 \cdot 0.5 + 0.1 \cdot 0.5 = 0.5$$

$$P(\alpha | R_1) = 0.9 \cdot \frac{0.5}{0.5} = 0.9$$

CALCOLO COMBINATORIO E PROBABILITÀ



Le conclusioni del nostro esperimento dipendono dalla conoscenza a priori del processo in studio?

Con che probabilità le particelle che interagiranno sul rivelatore 1 sono alpha?

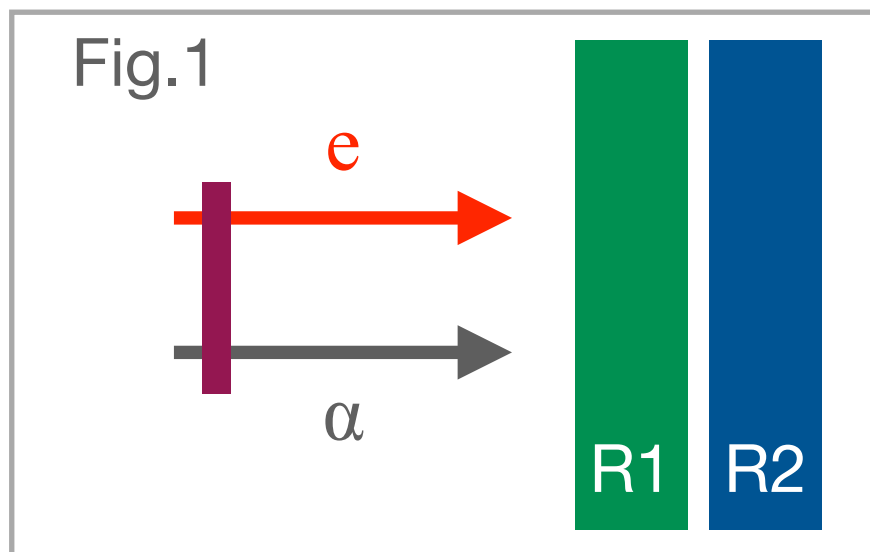
$$P(\alpha | R_1) = P(R_1 | \alpha) P(\alpha) / P(R_1) \quad \text{Probabilità a priori non uniforme 80-20}$$

$$P(R_1) = P(R_1 | \alpha) P(\alpha) + P(R_1 | e) P(e) = 0.9 \cdot 0.8 + 0.1 \cdot 0.2 = 0.74$$

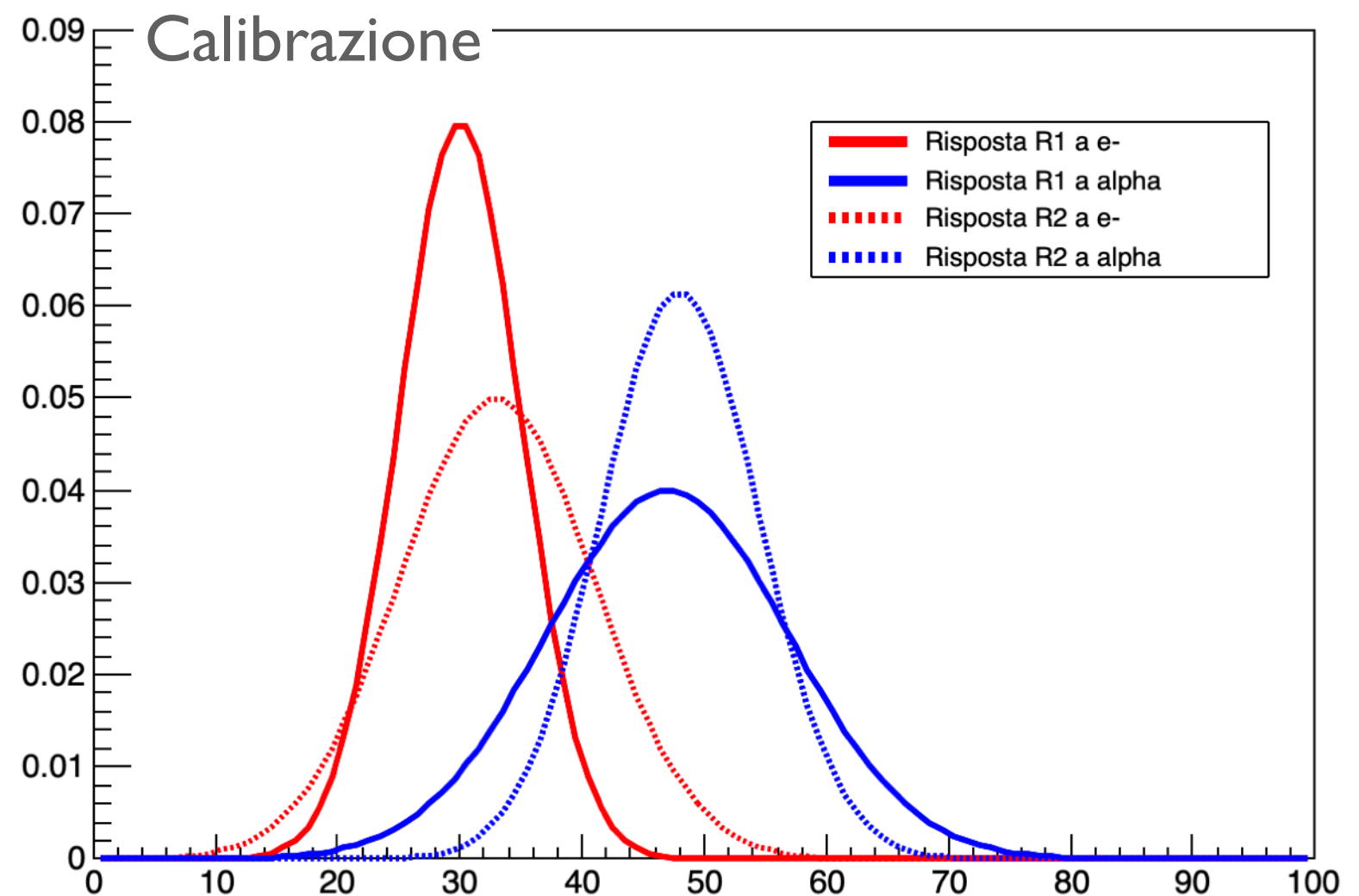
$$P(\alpha | R_1) = 0.9 \cdot \frac{0.8}{0.74} = 0.97$$

CALCOLO COMBINATORIO E PROBABILITÀ

- Un altro esempio: consideriamo un apparato sperimentale costituito da due rivelatori consecutivi che rivelano l'interazione di particelle (fig.1). Sull'apparato sperimentale arrivano 2 tipi di particelle: elettroni e particelle alpha (α).
- I due rivelatori hanno una diversa "risposta" ***che immaginiamo continua*** alle due tipologie di particelle posso selezionare con un certo livello di purezza un solo tipo di particelle ?



Precedentemente avrò calibrato il mio rivelatore su fasci di particelle di soli elettroni e sole α per determinarne la risposta di R1 e R2 ai diversi tipi di particella (fig.2)

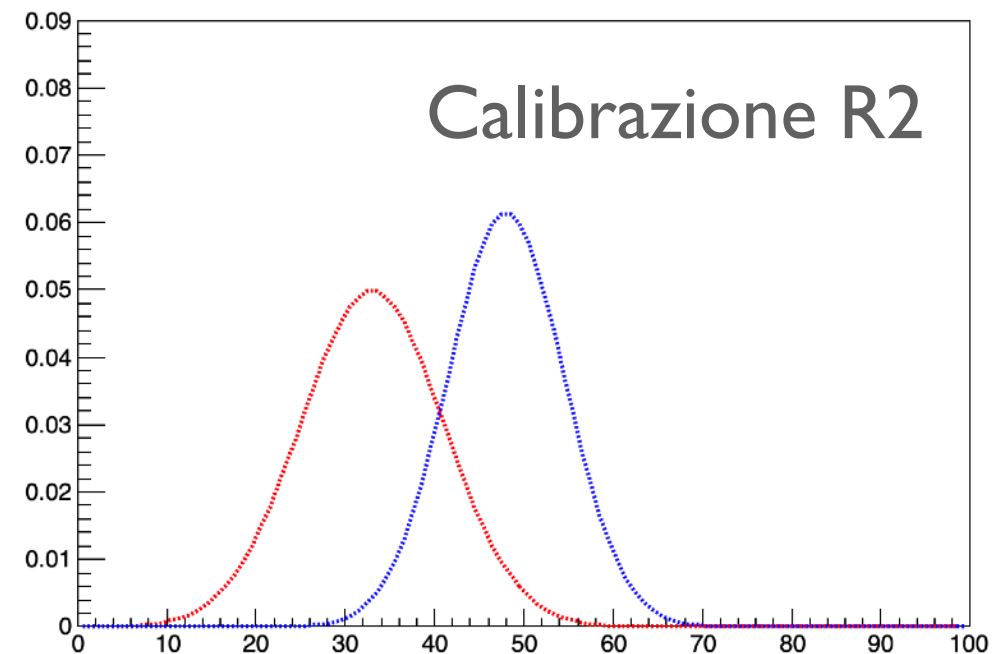
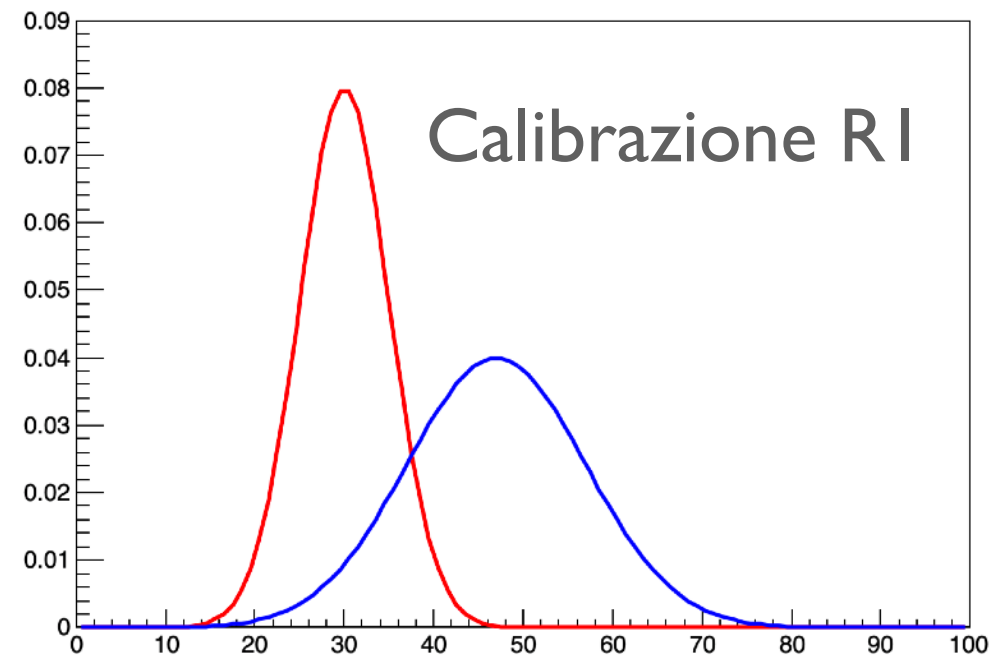


CALCOLO COMBINATORIO E PROBABILITÀ

```
root [73] for (int i=0; i<3; ++i) cout<<" Risposta R1 a e- P("<<i<<") = "<<fR1e.GetParameter(i)<<std::endl;
```

```
Risposta R1 a e- P(0) = 1  
Risposta R1 a e- P(1) = 30  
Risposta R1 a e- P(2) = 5
```

```
Risposta R1 a alpha P(0) = 1  
Risposta R1 a alpha P(1) = 47  
Risposta R1 a alpha P(2) = 10
```

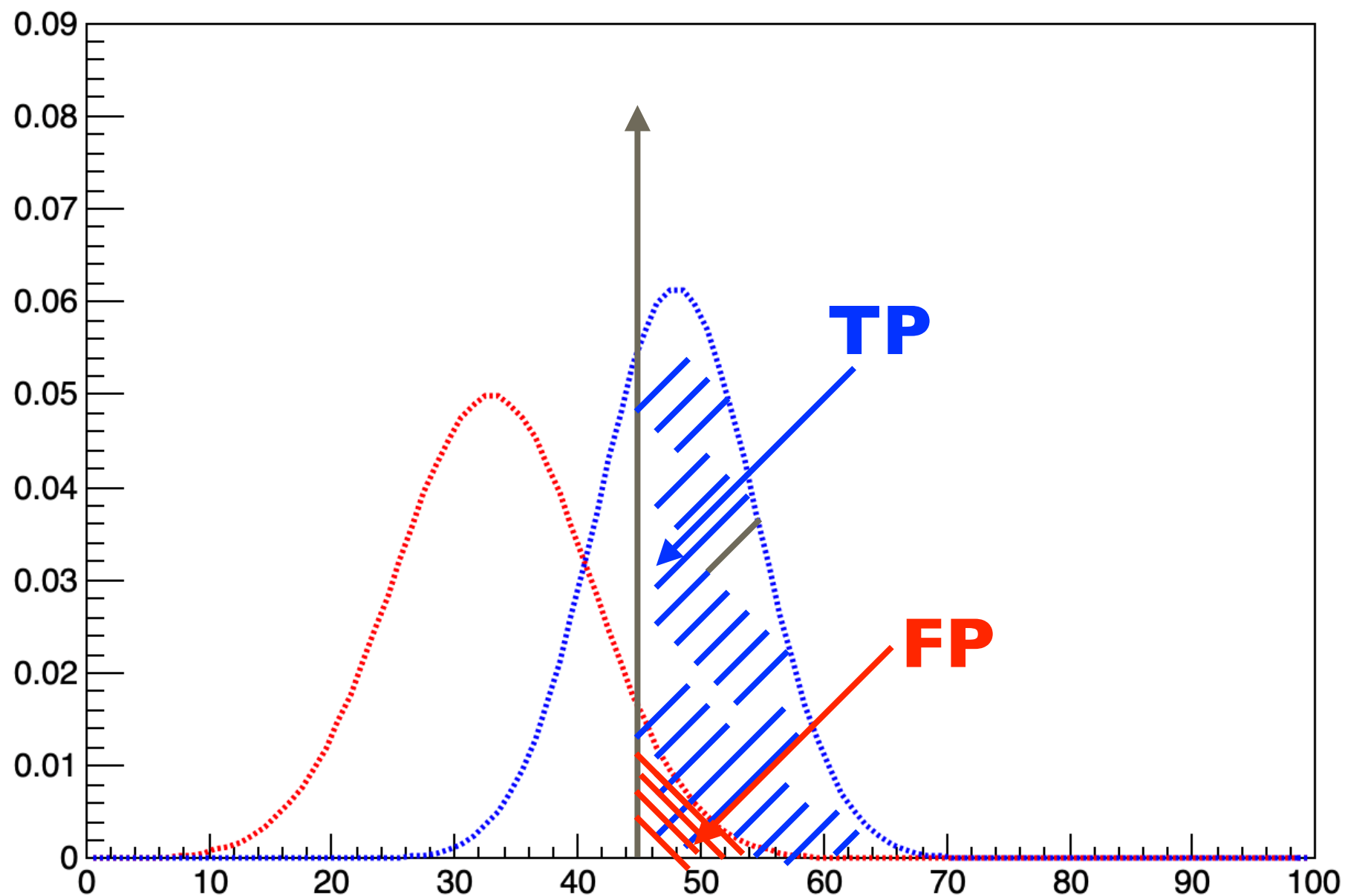


```
Risposta R2 a e- P(0) = 1  
Risposta R2 a e- P(1) = 33  
Risposta R2 a e- P(2) = 8
```

```
Risposta R2 a alpha P(0) = 1  
Risposta R2 a alpha P(1) = 48  
Risposta R2 a alpha P(2) = 6.5
```

R1 E R2 COME CLASSIFICATORI

- **Scelgo una soglia di discriminazione Thr : etichetto alpha una particella con risposta $> \text{Thr}$**
 - area tratteggiata in blu $\rightarrow$ eff per alpha (**TP**),
 - frazione di elettroni che passano la soglia di discriminazione per alpha (**FP**) in rosso



R1 E R2 COME CLASSIFICATORI

- **Scelgo una soglia di discriminazione Thr :**
 - **etichetto alpha una particella con risposta $> \text{Thr}$**
- **ESERCIZIO:**
 - costruire ROC e misurare AUC per R1 e per R2

FINE
