# Relazione Laboratorio 2: Distribuzione dei tempi di attesa per la rivelazione di raggi cosmici

Alessandra Arcuri

Febbraio 2020

## 1 Introduzione teorica

I raggi cosmici (costituiti prevalentemente da particelle e nuclei carichi ad alta energia provenienti dallo spazio) costituiscono un tipico esempio di *evento raro*: la variabile casuale corrispondente alla rivelazione di una particella (eseguita in laboratorio attraverso i *rivelatori a scintillazione*,) segue quindi la distribuzione di probabilità di Poisson.
Inoltre è possibile assumere, vista la loro natura, che gli eventi in considerazione siano indipendenti.
Si suppone, quindi, che la funzione di distribuzione per i tempi di attesa sia la seguente:

$$f(t) = \frac{1}{\tau} e^{-t/\tau} \tag{1}$$

essa è anche detta **funzione di distribuzione di Erlang** (con k=1, ossia dà la probabilità che si verifichi un evento successivo ad un evento verificatosi nell'istante arbitrario corrispondente a t=0). Essa può essere ricavata anche come caso particolare della *funzione di distribuzione Gamma*, ponendo $\alpha = k$ e $\beta = 1$.
Lo scopo dell'esercitazione è quindi quello di stimare il parametro $\tau$ utilizzando **quattro diversi metodi**:

- media aritmetica;
- metodo dei minimi quadrati;
- metodo dei minimi quadrati lineare;
- metodo del Binned Maximum Likelihood.

e poi svolgere un **test d'ipotesi** sulla forma della distribuzione, partendo da un set di dati raccolti in laboratorio.

# 2 Stima dei parametri

La stima dei parametri è quella parte dell'inferenza statistica che permette di ottenere da un certo set di parametri (ossia il campione di variabili casuali) delle certe stime puntuali, ossia dei valori numerici che vengono assunti come misure dei parametri (con l'opportuna incertezza associata).
Nel caso in esame, pur avendo ipotizzato la forma della funzione di distribuzione dei tempi di attesa, non è noto il valore del parametro $\tau$, che va quindi stimato a partire dal campione. Quest'ultimo è costituito da n misure di tempi di oscillazione $t_1, t_2, ..., t_n$ (con n = 171), presi tra l'istante arbitrario e l'istante in cui si verifica l'evento.
Vengono ora illustrati i quattro diversi metodi utilizzati per la stima del parametro $\tau$ e della sua relativa incertezza $\sigma_\tau$.

### 2.0.1 Media aritmetica

La media aritmetica del campione è uno stimatore che può essere ricavato attraverso il metodo dell' *Unbinned Maximum Likelihood*.
La **funzione di Likelihood** è definita come densità di probabilità congiunta delle v.c. indipendenti che costituiscono il campione, ossia i tempi di attesa, che in questo caso si ipotizza che abbiano funzione di distribuzione di Erlang (cfr. **(1)**). Si ottiene quindi la seguente funzione di distribuzione congiunta:

$$\mathcal{L}(\vec{t}, \tau) = \frac{1}{\tau} \prod_{i=1}^{n} e^{-t_i/\tau} = \frac{1}{\tau^n} \cdot e^{-\sum_i t_i/\tau} \tag{2}$$

Per il metodo dell'Unbinned Maximum Likelihood, gli stimatori dei parametri sono quelle funzioni del campione che **massimizzano** il valore della funzione di Likelihood. Per comodità si usa massimizzare il logaritmo della funzione di verosimiglianza, ossia la *log-verosimiglianza*:

$$\ln \mathcal{L} = -n \ln \tau - \frac{1}{\tau} \sum_{i=1}^{n} t_i \tag{3}$$

Lo stimatore di $\tau$ sarà quindi la funzione del campione tale che:

$$\frac{\partial}{\partial \tau} (\ln \mathcal{L}) = 0 \quad \Rightarrow \quad -\frac{n}{\tau} + \frac{1}{\tau^2} n\bar{t} = 0 \tag{4}$$

Da cui si ricava la stima finale di $\hat{\tau}$ che, come anticipato, corrisponde esattamente alla media aritmetica delle componenti del vettore che costituisce il campione:

$$\hat{\tau} = \bar{t} = \frac{1}{n} \sum_{i=1}^{n} t_i \tag{5}$$

Per quanto riguarda l'incertezza associata a $\hat{\tau}$, essa può essere calcolata sia utilizzando la *legge di propagazione della varianza*, ricordando che la media aritmetica segue la distribuzione gaussiana e considerando che la varianza associata alla distribuzione di Erlang è $\sigma^2_{Er} = \tau^2$, sia utilizzando la ***disuguaglianza di Cramer-Rao-Frechet***:

$$\sigma^2_{\hat{\tau}} = \frac{\sigma^2_{Er}}{n} = \frac{\tau^2}{n} \quad \Rightarrow \quad \sigma_{\hat{\tau}} = \frac{\bar{t}}{\sqrt{n}} = \frac{\hat{\tau}}{\sqrt{n}} \tag{6}$$

Si ottengono, quindi, i seguenti valori stimati:

$$\boxed{\hat{\tau} = 0.059s \quad \sigma_{\hat{\tau}} = 0.004s}$$

### 2.0.2 Metodo dei minimi quadrati (MMQ)

Il *Metodo dei minimi quadrati* afferma che la stima migliore di un parametro in base ad un dato campione sono quelle che minimizzano la forma quadratica:

$$Q^2 = \sum_{i=1}^{n} \frac{t_i - \mu_i}{\sigma_i^2} \tag{7}$$

Nel caso in esame, il MMQ è stato utilizzato a partire dall'istogramma costruito con il campione, utilizzando la seguente formula:

$$Q^2_{isto} = \sum_{i=0}^{n} \frac{n_i - \mu_i}{\sigma_i^2} \tag{8}$$

dove:

- $n_i$ è il numero di eventi nell'i-esimo bin;
- $\mu_i = E[n_i]$ è il valore di aspettazione dell'i-esimo bin; può essere ricavato considerando la distribuzione binomiale degli eventi nel bin. Se $\mathbf{n}$ è il numero totale di eventi, si ha quindi che $\mu_i = E[n_i] = nP_i = n\frac{1}{\tau}e^{-t_i^*/\tau} \cdot \Delta$, dove $t_i^*$ è il valore medio del singolo bin e $\Delta$ è l'ampiezza del bin.
- $\sigma^2$ è la varianza associata al numero di eventi nel bin; essa complica il calcolo analitico della derivata, quindi spesso si sostituisce ad essa il numero di eventi nel bin, ottenendo quindi la formula del *Metodo dei Minimi Quadrati semplificato*.

La formula semplificata risulta essere quindi la seguente:

$$Q^2_{isto} = \sum_{i=0}^{n} \frac{n_i - \mu_i}{n_i} \tag{9}$$

Per minimizzare la forma di quadratica è stato utilizzato il **metodo grafico** mediante un programma in grado di calcolare il vertice della parabola, corrispondente quindi alla migliore stima di $\hat{\tau}$.

Per calcolare l'incertezza corrispondente, è stata graficata la retta corrispondente a $Q^2 + 1$ che, intersecandosi con la curva del $Q^2$, permette di determinare l'intervallo di semiampiezza pari a una deviazione standard. La curva totale ottenuta è la seguente:

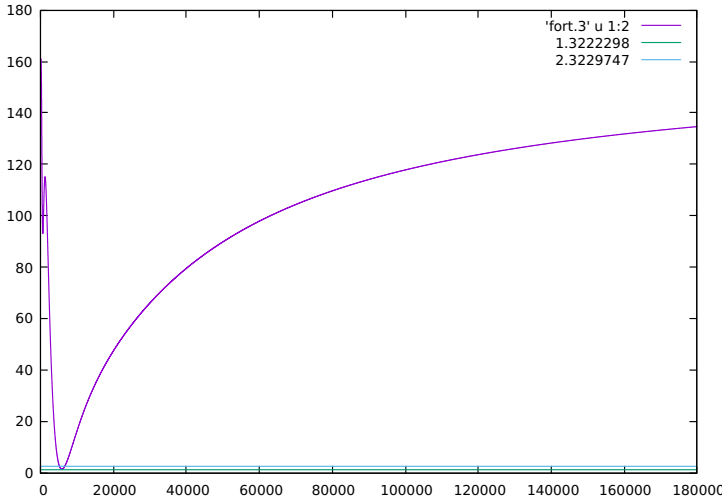


Figure 1: MMQ

Ingrandendo sulla porzione che segue andamento parabolico:

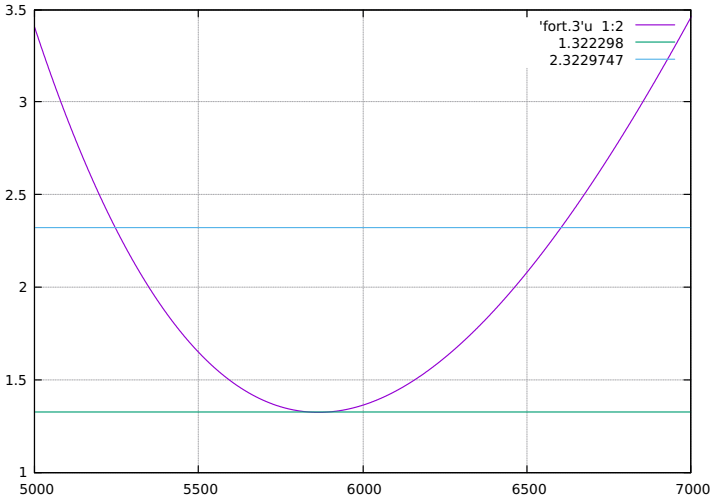


Figure 2: Parabola MMQ

I valori ottenuti sono quindi i seguenti:

$$\boxed{\hat{\tau} = 0.059s \quad \sigma_{\hat{\tau}} = 0.006s}$$

### 2.0.3 Metodo dei minimi quadrati lineare (MMQL)

Il caso precedente può essere *linearizzato*, prendendo al posto di $n_i$ il suo logaritmo: $y_i = ln(n_i)$. Il valore di aspettazione di $y_i$ diventa quindi esprimibile attraverso una relazione lineare: $E[y_i] = \theta_1 + \theta_2 t_i^*$ dove:

- $\theta_1 = ln(\tau) + ln(n) + ln(\Delta)$
- $\theta_2 = -1/\tau$

In questo modo, calcolando l'inverso del coefficiente angolare della retta, è possibile ottenere il valore stimato del parametro cercato.
E' possibile trovare il valore del coefficiente angolare minimizzando la funzione di $Q^2$ modificata con i nuovi parametri, ottenendo quindi:

$$\hat{m} = \frac{S_{00}S_{11} - S_{10}S_{01}}{S_{00}S_{20} - S_{10}^2} \tag{10}$$

dove è stata utilizzata la notazione:

$$S_{00} = \sum_{i=1}^{n} \frac{1}{\sigma_i^2}, \quad S_{01} = \sum_{i=1}^{n} \frac{y_i}{\sigma_i^2}, \quad S_{10} = \sum_{i=1}^{n} \frac{t_i^*}{\sigma_i^2}, \quad S_{11} = \sum_{i=1}^{n} \frac{t_i^* y_i}{\sigma_i^2}, \quad S_{20} = \sum_{i=1}^{n} \frac{(t_i^*)^2}{\sigma_i^2}$$

Per quanto riguarda l'incertezza su $\tau$, è stata prima calcolata l'incertezza sulle $y_i$ (ossia $\sigma_i$) attraverso la legge di propagazione della varianza, la quale è stata utilizzata poi per ricavare $\sigma_{\hat{\tau}}$.
I valori ricercati risultano essere:

$$\boxed{\hat{\tau} = 0.059s \quad \sigma_{\hat{\tau}} = 0.008s}$$

### 2.0.4 Metodo del Binned Maximum Likelihood (MML)

Il *metodo del Binned Maximum Likelihood*, come suggerito dal nome, segue lo stesso principio dell'Unbinned Maximum Likelihood: in questo caso si parte, però, dal campione relativo ai diversi intervalli dell'istogramma. Applicando la funzione di Likelihood a questo nuovo campione, si ottiene quindi una **distribuzione multinomiale**:

$$\mathcal{L}(n_1, ..., n_m; \tau) = \frac{N!}{n_1!...n_m!} P_1^{n_1} \cdots P_m^{n_m} \tag{11}$$

dove:

- $n_i$ è il numero di misure che ricadono nell'intervallo e $\sum_{i=1}^{m} n_i = N$;
- $P_i$ è la probabilità che la misura cada nell'i-esimo bin.

Minimizzando, quindi, la funzione di log-verosimiglianza, si ottiene ancora una volta il valore del parametro $\tau$. Per ricavare l'intervallo di confidenza di semi-ampiezza pari ad una deviazione standard e quindi l'incertezza, è necessario intersecare la curva corrispondente a $ln(\mathcal{L})$ con la retta data da $ln(\mathcal{L}_{max}) - \frac{1}{2}$. Si ottiene così il seguente grafico:

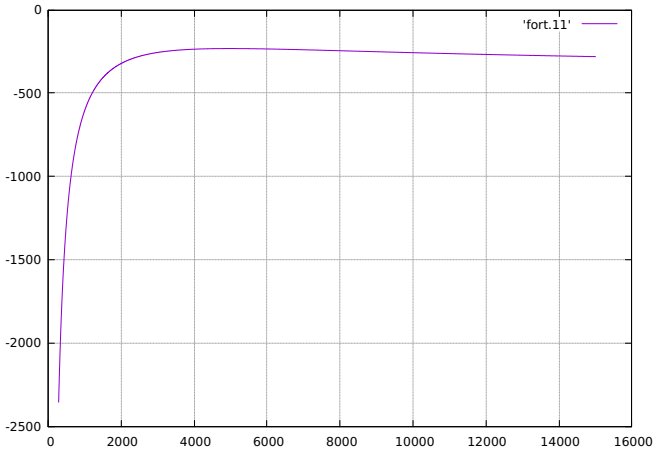


Figure 3: MML

Ingrandendo sulla porzione che segue andamento parabolico:

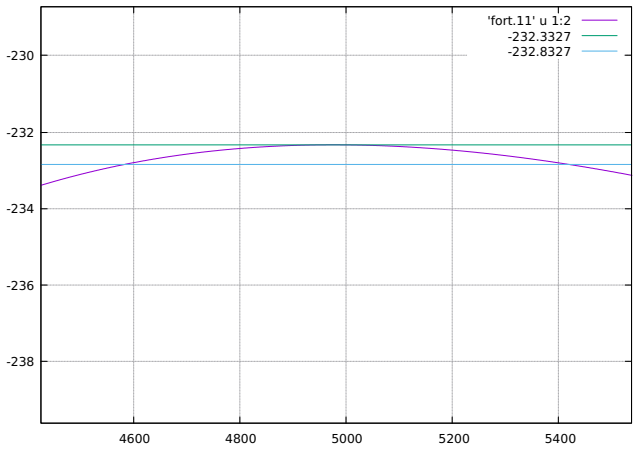


Figure 4: Parabola MML

I valori stimati sono quindi i seguenti:

$$\hat{\tau} = 0.050s \quad \sigma_{\hat{\tau}} = 0.004s$$

# 3 Test d'ipotesi

Il ***test d'ipotesi*** è quella parte dell'inferenza statistica che permette di valutare quantitativamente l'accordo tra un modello statistico (ossia l'*ipotesi nulla* $H_0$, relativa alla forma della funzione di distribuzione di un set di variabili casuali) e le osservazioni sperimentali. Se $x_1, ..., x_n$ è il campione di partenza, si introduce una funzione del campione $t(x_1, ..., x_n)$, detta **statistica di test**, di cui si conosce la funzione di distribuzione $\Phi_0(t)$ secondo l'ipotesi nulla $H_0$. Dopo aver stabilito il *livello di significatività* $\alpha$, il quale indica la probabilità di rigettare l'ipotesi nulla anche se è corretta, si definisce:

$$\alpha = \int_{t_\alpha}^{t_{max}} \Phi_0(t)dt \tag{12}$$

dove $t_\alpha$ è un valore tabulato in base alla funzione $\Phi_0(t)$ e $0 < \alpha < 1$; diminuendo $\alpha$, diminuisce anche la probabilità di commettere l'errore di rigettare l'ipotesi nulla anche se vera.
Se $\tilde{t}$ è il valore ottenuto dalla statistica di test, si rigetta l'ipotesi nulla nel caso in cui $\tilde{t} > t_\alpha$.

## 3.1 Test di $\chi^2$

Nel caso in esame, si è fatto uso del *test di* $\chi^2$: esso utilizza come statistica di test una funzione che ha distribuzione di Chi quadro (quindi $\Phi_0(t) \equiv \chi^2$):

$$\chi^2 = \sum_{i=1}^{n} \frac{(n_i - \mu_i)^2}{\sigma_i^2} \tag{13}$$

Applicando la formula al campione costituito dal numero di misure in ogni intervallo dell'istogramma delle misure (i.e. $n_1, ..., n_m$, con $m$ il numero di bin dell'istogramma) si ottiene:

$$\chi^2 = \sum_{i=1}^{m} \frac{(n_i - E[n_i])^2}{E[n_i]} = t(n_1, ..., n_m) \tag{14}$$

Dove $E[n_i]$ è il valore di aspettazione di ogni bin, calcolato utilizzando una delle quattro stime di $\tau$ ottenute:

$$\mu_i = E[n_i] = n \cdot p_i \approx n \cdot f(t_k; \hat{\tau})\Delta \tag{15}$$

L'ipotesi nulla $H_0$ è quindi, in questo caso, che la distribuzione $f(t_k; \hat{\tau})$ sia quella di Erlang, utilizzata per calcolare $E[n_i]$.
Per quanto riguarda i valori ottenuti, bisogna tenere conto del fatto che, negli ultimi 5 bin dell'istogramma, $E[n_i]$ risultava essere $< 5$; si è scelto quindi di effettuare il calcolo della **(14)** utilizzando solo i primi cinque bin dell'istogramma.

I valori di $\tilde{t}$ ottenuti vengono confrontati con i valori di $t_\alpha$ associati a $\chi^2_3$, cioè di Chi Quadro a $\nu = 3$ gradi di libertà (5 bin - 1 parametro stimato ($\tau$) -1). La tabella a cui si è fatto riferimento è la seguente:

| $\nu \vert \alpha$ | 0.1 | 0.05 | 0.025 |
|---|---|---|---|
| 1) | 2.706 | 3.841 | 5.024 |
| 2) | 4.605 | 5.991 | 7.378 |
| 3) | 6.251 | 7.815 | 9.348 |
| 4) | 7.779 | 9.488 | 11.143 |

Table 1: Tabella test di Chi Quadro

Mentre i valori di $\tilde{t}$ corrispondenti alle quattro stime $\hat{\tau}$:

| $\hat{\tau}$ | $\tilde{t}$ |
|---|---|
| $\hat{\tau}_1$ | 2.617 |
| $\hat{\tau}_2$ | 2.564 |
| $\hat{\tau}_3$ | 2.603 |
| $\hat{\tau}_4$ | 4.259 |

Table 2: Valori ottenuti dal test

L'ipotesi nulla ($H_0$: i tempi di attesa seguono la distribuzione di Erlang) non viene quindi rigettata per valori di $\alpha$ pari a 0.1, 0.05, 0.0025 e $\nu = 3$.

# 4 Istogramma complessivo e commenti

Per confrontare visivamente i quattro metodi di stima utilizzati è possibile costruire l'istogramma complessivo delle misure e sovrapporre ad esso le quattro funzioni di Erlang corrispondenti:

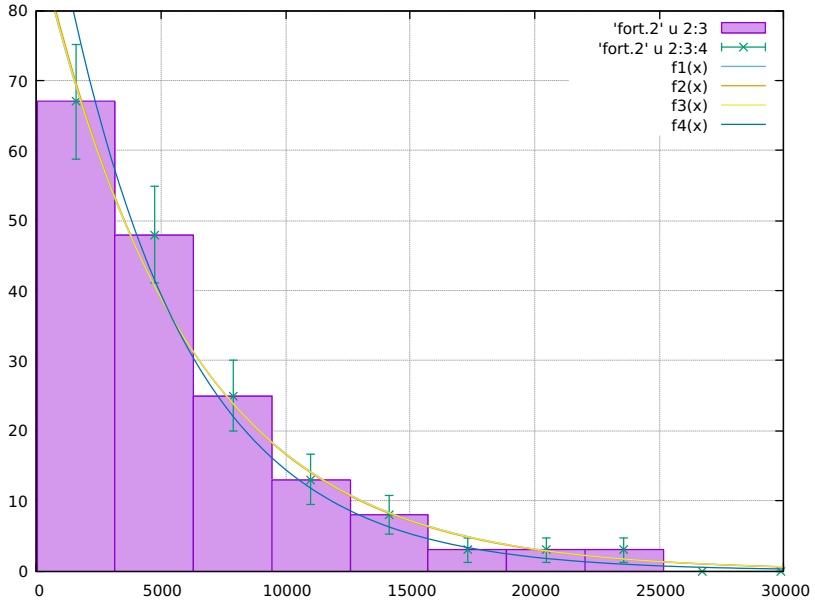


Figure 5: Istogramma complessivo con i quattro metodi

Come si può evincere dalla figura, tutte e quattro le distribuzioni risultano rientrare nell'errore corrispondente ad ogni bin dell'istogramma.
Per quanto riguarda la compatibilità tra i diversi $\hat{\tau}_i$, essa è confermata per i primi tre, mentre $\hat{\tau}_4$ (ossia quello ricavato da MML) non risulta essere compatibile con gli altri valori. Ciò potrebbe essere migliorato aumentando la quantità statistica a disposizione, ossia ripetendo l'esperimento o ancora eseguendo delle simulazioni di esso.