In [1]:

```
#la legge dei grandi numeri e il teorema del limite centrale ci permettono di dire che, d
ata una certa popolazione, e presi
#x campioni, più il numero dei campioni è grande preso dalla popolazione più la media del
le medie dei campioni approccerà
#quella della popolazione, e si distribuiranno lungo una gaussiana. Se ad esempio da una
popolazione di centomila prendo 500
#campioni da 10 e 500 campioni da 100, le due gaussiane che si formeranno avranno una med
ia simile, solo che quella con
#sample size più grande avrà una deviazione standard minore attorno alla sua media.
#Poi, quando prendiamo un campione da una popolazione di cui non sappiamo la distribuzion
e, noi possiamo comunque dire che
#le medie campionarie si distribuiranno lungo una gaussiana.
#Ora, se prendo un campione di 50 da una popolazione di 1 milione e ne faccio la media, d
ove non conosco né la media né la
#deviazione standard della popolazione, posso comunque usare una distribuzione di t di st
udent attorno alla media del mio
#campione, che ha code più larghe, e quindi un intervallo di confidenza al 95% rispetto a
una curva normale avrà i limiti
#dell'intervallo più ampi. Mettiamo caso la media del campione è 100 e i limiti dell'inte
rvallo sono stati 98.5 - 101.5 al
#95% di confidenza. Io in realtà sto dicendo che, se estraggo altri campioni con stessa g
randezza dalla stessa popolazione
#e ne calcolo gli intervalli di confidenza al 95%, se i campioni estratti sono 100, 95 co
nterranno nel loro intervallo di
#confidenza la vera media della popolazione, mentre 5 di questi non la conterranno. Quind
i, se sto facendo un lavoro e uso
#solo un campione e ne faccio l'intervallo di confidenza al 95%, usando solo questo campi
one per orientarmi su dove si trovi
#la popolazione al 95%, io sto assumendo che il mio campione fa parte di quei 95 che cont
engono la media della popolazione
#sui 100 che potrei estrarre.

#nota bene: il teorema vale anche se la popolazione non è normalmente distribuita. le med
ie dei samples saranno cmq normalmente
#distribuite, e approcceranno sempre più la normal distribution più è grande la sample si
ze.
```

In [2]:

```
#ex2:

#L'azienda ABC produce viti. i valori della lunghezza delle viti seguono una normale, con
population std di 2 mm.
#basandoti su un sample di 50, costruisci un intervallo di confidenza del 90% per la medi
a (lunghezza viti) della popolazione.
#Assumiamo tuttavia di non conoscere né la media né la std della popolazione.
```

In [6]:

```
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as stats
```

In [3]:

```
#creo popolazione di 1 milione di viti che si distribuiscono lungo una normale.

mean = 100
std = 2
size_abc = 1000000 #della population totale
```

In [7]:

```
np.random.seed(4)
pop = np.random.normal(loc = mean, scale = std, size = size_abc)
```

In [10]:

```
#creo uno sample di 50 viti
sample_size = 50
```

In [11]:

```
np.random.seed(1)
sample = np.random.choice(pop, sample_size, replace = False)
```

In [12]:

```
sample
```

Out[12]:

```
array([102.59027032, 102.02657944,  99.26742988, 103.33867702,
        97.77169871,  99.73609551,  99.39014832,  99.92356623,
        97.92544099,  99.96827327, 102.60974374,  98.31166016,
        99.95277111, 101.9838262 , 102.20271966, 100.396764  ,
        95.82424373, 100.5471759 ,  99.59795894, 100.04724152,
       101.93541093,  98.62787698, 100.85313982,  98.55858236,
        98.86740259, 100.05160492,  99.94467685,  99.93949911,
        99.99155601, 102.51150372, 102.74321614, 101.64443194,
        99.55262934,  99.88919766, 102.01132637,  93.44675737,
        99.83199028, 104.46915574, 100.72023086, 101.76929456,
        96.47363041, 102.22715495, 103.54290521, 104.24655194,
       104.43264119,  98.74785217,  98.76798047, 100.87069417,
       102.44293457, 101.17225014])
```

In [13]:

```
#media del sample

sample.mean()
```

Out[13]:

```
100.47392726861327
```

In [14]:

```
#media pop

pop.mean()
```

Out[14]:

```
99.99440398643918
```

In [15]:

```
#per creare intervalli di conf per la population mean si ha bisogno della sample mean (po
int of
# estimate of the mean), la pop std /radice di size of the sample = std error. C'è un pro
blema: nella realtà è difficile
# che si abbia la std della pop. Possiamo trovare un point of estimate of pop std, che è
la sample std. Le stime
# (in questo caso sono due, quelli della media e std) sono meno precise, particolarmente
per small sample sizes
#quindi la normal distribution non è appropriata, e abbiamo bisogno di una distr più cons
ervativa, che ha maggiore probabilità
# nelle code
# questa è la student's t distribution. E visto che le code sono più larghe, anche gli in
tervalli di conf saranno più ampi
#a parità di livello di confidenza
```

In [ ]:

```
#il punto di stima della media è la media del sample, quello della std è la std del sampl
e
#(deviazione standard) con 1 gradi di libertà.
```

In [16]:

```
point_est_mean = sample.mean()
point_est_mean
```

Out[16]:

```
100.47392726861327
```

In [18]:

```
point_est_std = sample.std(ddof = 1) #cioè si usa la std del sample moltiplicato radice d
i(n/n-1)
point_est_std
```

Out[18]:

```
2.182620742091214
```

In [19]:

```
#pongo un livello di confidenza del 90%
conf = 0.90
```

In [22]:

```
#calcolo l'errore standard
standard_error = point_est_std / np.sqrt(sample_size)
standard_error
```

Out[22]:

```
0.30866918549822236
```

In [23]:

```
#calcolo intervalli di conf 90% nella t student devo mettere sempre un grado di libertà

stats.t.interval(conf, loc = point_est_mean, scale = standard_error, df = sample_size -
1 )
```

Out[23]:

```
(99.95642767035977, 100.99142686686676)
```

In [24]:

```
#spiegazione:
#Dato 1 milione di viti, non so la media né la std di questa popolazione, prendo un campi
one di 50 e calcolo la media, e uso
#quindi una t-student per questo sample,calcolo poi l'intervallo di confidenza e mi trovo
i due valori critici attorno la media
#di questo campione, e questo esperimento mi dice che se faccio lo stesso esperimento mol
te altre volte, cioè prendo altri
#campioni da 50 e calcolo il lvl di confidenza al 90%, il vero significato è che il 90% d
i questi campioni avrà la
#population mean nei loro valori critici attorno alla media del campione, sapendo che ogn
i campione avrà una distr di t student.
#se non so quindi la std della popolazione, allora la t student mi farà avere intervalli
di confidenza con valori critici
#più ampi dalla media, rispetto ai valori critici degli intervalli di confidenza della di
str normale di un campione su cui si sa
#la std della popolazione.
```

In [ ]:

```

```