In [ ]:

```
#L'azienda Y produce viti che devono essere lunghe 100 mm. La lunghezza delle viti segue una distr normale con una std
#(dev standard) della popolazione = 2.
#Le macchine che producono viti devono essere ricalibrate una volta a settimana, e si dovrà quindi controllare se la macchina
#è correttamente calibrata nel produrre viti da 100 mm di lunghezza. Basandoti su un sample di 20 viti, ottieni una lunghezza
#media di 100.929 mm. Testa con un lvl di 2% significatività se la macchina è correttamente calibrata o corrotta.
#Calcola il p-value e la z-statistic.

#L'ipotesi nulla H0 è che la lunghezza media prodotta dalla macchina sia 100, l'ipotesi alternativa H1 è che sia diversa da 100.
#Se non accetto l'ipotesi nulla, l'implicazione che la macchina sia corrotta e produca mediamente viti diverse da 100 è vera.
#Quindi supponendo che la lunghezza delle viti seguano una curva normale, uso una confidenza del 98% sulla curva normale
#con l'assunzione che la media interna all'intervallo sia quello dell'ipotesi nulla H0.

#Se la sample mean è fuori quest'intervallo, ho il 2% di probabilità, quindi una probabilità bassa che l'ipotesi nulla sia vera, quindi le macchine
#producano viti da 100 mm di lunghezza media, pur avendo una media del campione delle viti fuori dall'intervallo di confidenza.
#Quindi in questo caso, data la bassa probabilità del 2% di avere una sample mean fuori l'intervallo di confidenza del 98%
#pur avendo una ipotesi nulla vera, rifiuto l'ipotesi nulla. Se la sample mean fosse all'interno dell'intervallo di confidenza,
#ho il 2% di probabilità di commettere un errore, ovvero sia che la sample mean sia all'interno dell'intervallo di confidenza,
#anche se la macchina produce viti che hanno una media diversa da 100 mm.In questo caso non rifiuterei l'ipotesi nulla, e
#avrei una confidenza del 98% di non commettere un errore assumendo che la macchina produca viti di 100mm in media.
```

In [ ]:

```
#qual è quindi il  lvl di significatività migliore? l'errore di tipo 1, indica la prob di rifiutare l'ipotesi nulla quando
# è vera (la media di 100). Se l'intervallo di conf è al 90%, la prob di rifiutare l'ipotesi nulla è 10% se la sample mean
#è fuori l'intervallo. D'altra parte, diminuire il lvl di significatività, quindi aumentare il lvl di confidenza del tipo,
# avere diminuire il livello di significatività al 5%, aumentando dunque l'intervallo di conf al 95%, aumenta la prob
#dell'errore di tipo 2, cioè non rigettare l'ipotesi nulla quando è falsa ed è vera Ha.
#Cioè l'intervallo di conf più alto
#aumenta la prob di NON rigettare l'ipotesi nulla, quindi per rigettare l'ipotesi nulla in questo caso dobbiamo avere
#molti campioni estremi con medie estreme, perché l'intervallo
# è più largo, E quindi è più difficoltoso provare che la pop sia veramente di 100, perché noi effettivamente stiamo
# aumentando l'intervallo di confidenza e il range in cui potrebbe essere, quindi ho più probabilità di accettare H0.
#Quindi: se accetto H0, e H0 è vera, conclusione corretta. Se non rifiuto H0, ma è vera Ha, c'è l'errore di tipo 2.
#se rifiuto H0, e H0 è vera, errore di tipo 1. Se rifiuto H0, ma è vera Ha, conclusione corretta.
```

In [ ]:

```
#qual è il lvl di significatività migliore? dipende. C'è un trade off in questo caso, o produrre viti troppo lunghe/corte,
#o stoppare le macchine per ricalibrare (che può costare). Più è alto il lvl di significatività e quindi basso l'intervallo
#di conf, più si sta aumentando la prob di rigettare l'ipotesi nulla e quindi di ricalibrare le macchine, perché il sample mean
```

```
#sarebbe fuori l'intervallo di conf quindi dirai che la pop mean sarà diversa da 100. D'a
ltra parte, più abbassi il lvl di
#significatività e dunque aumentando il lvl di confidenza, più si sta abbassando la prob
di rigettare l'ipotesi nulla e
#quindi aumentando la prob di avere viti diverse da 100, perché in questo caso l'interval
lo di confidenza sarebbe più ampio,
#e per rifiutare l'ipotesi nulla e dunque accettare l'ipotesi che il macchinario sia corr
otto e non produca viti da 100mm
#in media, occorrerebbe avere dei campioni che abbiano medie estreme che ricadano fuori l
'intervallo di confidenza più largo,
#pure nell'occasione che la macchina produca effettivamente viti troppo lunghe/corte.
```

In [1]:

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

In [2]:

```
#la macchina corrotta produce viti con

mu = 100.7 #unknown
pop_std = 2 #known
```

In [3]:

```
sample_size = 20
#sappiamo che le viti sono normalmente distribuite. Se quindi la pop è normalmente distri
buita, allora non importa la sample
#size anche ogni campione sarà normalmente distribuito.
```

In [4]:

```
#creo il sample di 20 normalmente distribuito

np.random.seed(123)
sample = np.random.normal(loc = mu, scale = pop_std, size = sample_size)

sample
```

Out[4]:

```
array([ 98.52873879, 102.69469089, 101.265957  ,  97.68741057,
        99.5427995 , 104.00287307,  95.84664151,  99.84217474,
       103.23187252,  98.9665192 ,  99.3422277 , 100.51058206,
       103.68277925,  99.42219601,  99.81203608,  99.83129745,
       105.11186017, 105.07357218, 102.7081078 , 101.4723728 ])
```

In [5]:

```
point_est_mean = sample.mean()
point_est_mean
#il punto di stima della media della popolazione è uguale alla media del campione
```

Out[5]:

```
100.92883546391059
```

In [6]:

```
std_error = pop_std / np.sqrt(sample_size)
std_error
```

Out[6]:

```
0.4472135954999579
```

In [7]:

```
H0 = 100
#pongo l'ipotesi nulla uguale a 100, ovvero che il macchinario produca viti da 100mm in m
```

```
edia.
# H0 -> ipotesi nulla = 100
# Ha != 100 -> cioè la vera media è diversa da 100.
```

In [8]:

```
conf = 0.90
#scelgo una confidenza del 90%
```

In [9]:

```
#creo l'intervallo di conf al 90%

stats.norm.interval(conf, loc = H0, scale = std_error)

# sto quindi assumendo che la vera media sia 100
#vediamo che però la sample mean è fuori questo intervallo
```

Out[9]:

```
(99.26439909541989, 100.73560090458011)
```

In [ ]:

```
#la sample mean è fuori l'intervallo, quindi è molto probabile che l'ipotesi nulla sia da
rigettare.
#anche se sappiamo per ipotesi, anche non conosciuta, il macchinario sia effettivamente c
orrotto e produca viti di 100.7 mm.
```

In [10]:

```
x = np.linspace(96, 104, 1000) #metto mille valori tra 96 e 104
y = stats.norm.pdf(x, loc = H0, scale = std_error)
```

In [16]:

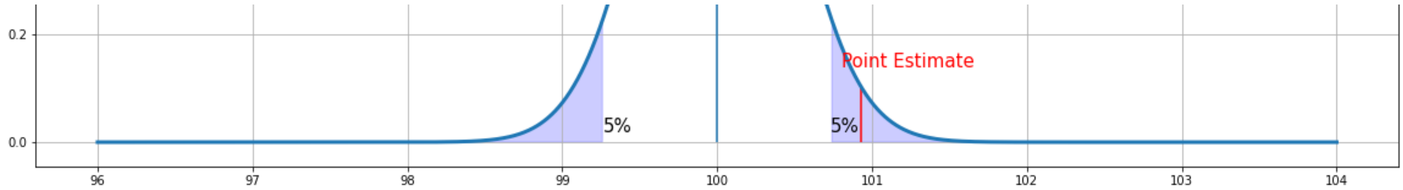
```
left, right = stats.norm.interval(conf, loc = H0, scale = std_error)
```

In [19]:

```
plt.figure(figsize = (20, 8))
plt.plot(x, y, linewidth = 3, label = "Normal Distribution")
plt.vlines(x = H0, ymin = 0, ymax = 0.90)
plt.vlines(x = point_est_mean, ymin = 0, ymax = 0.1, color = "red")
plt.fill_between(x, y, where = ((x > right) | (x < left)), color = "blue", alpha = 0.2)
plt.annotate("5%", xy = (100.73, 0.02), fontsize = 15)
plt.annotate("5%", xy = (99.26, 0.02), fontsize = 15)
plt.annotate('Reject H0 !\nMachine is corrupted!', xy = (96, 0.5), fontsize = 30)
plt.annotate("H0: Mean = 100", xy = (100.05, 0.3), fontsize = 15)
plt.annotate("Point Estimate", xy = (100.8, 0.14), color = "red", fontsize = 15)
plt.grid()
plt.title("Z-test two sided", fontsize = 20)
plt.ylabel("pdf", fontsize = 15)
plt.legend(fontsize = 15)
```

Out[19]:

```
<matplotlib.legend.Legend at 0x24785435d60>
```



In [20]:

```
#il point of estimate è fuori l'intervallo di conf, quindi si rigetta l'ipotesi nulla e l
a macchina è corrotta perché non
#produce viti (popolazione) di media 100
#Nota bene: usiamo distr normale perché sappiamo la varianza della popolazione, altriment
i la distribuzione da usare
# è una t-student
```

In [21]:

```
#Dato un livello di confidenza del 90%, il punto di stima della media è fuori l'intervall
o di confidenza con media 100, cioè
# abbiamo il 10% di probabilità di avere dei valori fuori l'intervallo di confidenza, dat
o che comunque il macchinario produca
# effettivamente viti da 100 mm in media.
#Se aumentassi il lvl di confidenza, ad esempio al 98%, la media del sample ricadrebbe ne
ll'intervallo di confidenza, pur avendo
#una macchina che produce viti con una media diversa da 100 mm, dunque corrotta.
```

In [22]:

```
#Calcolo il p-value

#spiegazione teorica p-value: è la probabilità che il possibile rifiuto dell'ipotesi null
a sia solo dovuto al caso.
#poniamo il caso di voler verificare che il reddito medio delle famiglie di milano sia si
gnificativamente diverso da 1500 euro.
#poniamo il caso di aver estratto un campione che casualmente è constituito dalle famigli
e più ricche di milano. In questo caso
#avremmo che il campione avrebbe un reddito medio significativamente maggiore di 1500 eur
o, e dunque cadrebbe fuori
#l'intervallo di confidenza, dunque rifiuteremo l'ipotesi nulla. Ma questo rifiuto è solo
dovuto al caso, in quanto
#l'universo è costituito sia da famiglie povere che medie che ricche, ma casualmente il m
io campione ha pescato solo famiglie
#ricche. Quindi il mio p value mi dice che il possibile rifiuto dell'ipotesi nulla sia do
vuto alla casualità del campione
#che ho estratto che non rispecchia la situazione dell'universo.

#Calcolo p value in pratica: è la probabilità di osservare valori più estremi della stati
stica test osservata.
#Poniamo caso nel nostro test d'ipotesi abbiamo la media del campione che cade nella regi
one di non rifiuto, e il valore
#essendo un test unilaterale sx è z = -1.68. Quindi valori più estremi significa valori i
nferiori a -1.68.
#Quindi il p value sarebbe l'area che va da meno infito a -1.68.In questo caso sarebbe il
p value sarebbe 0.0465.
#Se il p value è > del lvl di significatività, allora non rifiuto l'ipotesi nulla, altrim
enti rifiuto.
#Se il lvl di significatività è 1%, cioè 0.01 con un intervallo di confidenza del 99%, al
lora rifiuto l'ipotesi nulla.
#se vi è una Bilaterale, allora il p value si calcola come l'area a dx del point of estim
ate, moltiplicato a due, dove
#il point of estimate per calcolarne l'area si prende sempre negativo, e poi moltiplico d
ue.
```

In [23]:

```
z_stat = (point_est_mean - H0)/ std_error
z_stat
#calcolo la z-statistic, che ci dice in questo caso che la media del sample è 2 std_error
lontano dalla media
```

Out[23]:

2.0769392372166267

In [24]:

```
stats.norm.cdf(-abs(z_stat))
#calcola la prob della coda. Se metto meno, calcola la prob della coda a sx.
#sia che stia facendo una unilaterale a sx o dx, devo calcolare sempre il z_stat col segn
o meno, perché mi serve l'area
#che va dalla z_stat a infinito, sia esso che sia dx che sx, se metto col valore positivo
, anche in caso di unilaterale dx,
#io starei calcolando l'area che va da - infinito alla z-statistic a destra della media d
ell'ipotesi nulla.
```

Out[24]:

0.01890358386435081

In [25]:

```
p_value = 2 * stats.norm.cdf(-abs(z_stat))
p_value
# è la p delle due p sommate delle code. Lo faccio perché in questo caso sto facendo una
bilaterale.
```

Out[25]:

0.03780716772870162

In [26]:

```
#siccome il p-value < lvl di significatività, cioè 0.037 < 0.1, rifiuto l'ipotesi nulla.
```

In [ ]:

```

```