DISTRIBUCIONES SET DE DATOS
(13 de dic. de 2020)

<u>dataSet01</u>

"Este set de datos contiene 10.000 edades extraídas de una base de datos a la cual se quiere realiza un UPDATE (actualización de datos), aumentando las edades en 15 unidades, sin embargo el parámetro de la base de datos es de tipo NUMERIC de 2 dígitos, por lo tanto, todas aquellas edades mayores de 85 lanzaran un error a la base de datos. Calcule la probabilidad que de 10 edades seleccionadas, ninguna marque error."

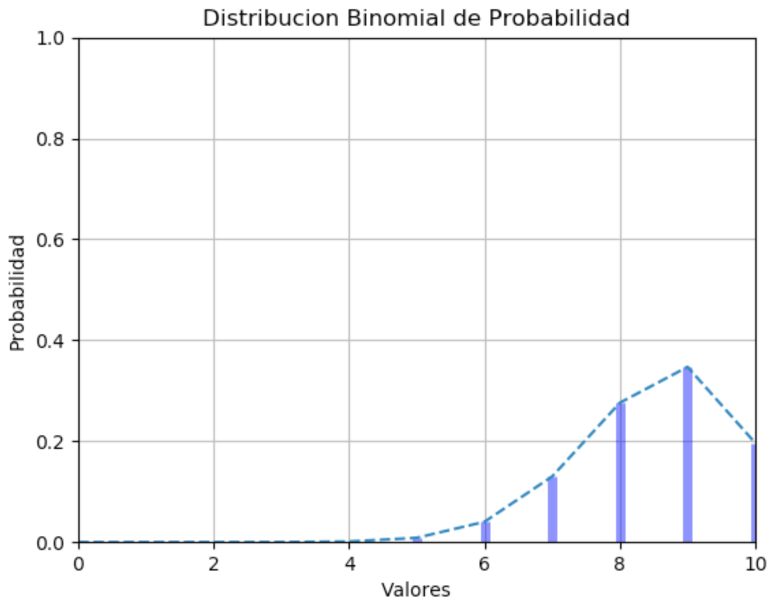
Al tratarse de una variable discreta, es posible plantear el set de datos mediante una distribución binomial, en donde se considera como éxito que la edad seleccionada no marque error (sea menor de 85), y como fallo que marque error (se mayor o igual a 85). Se obtendría los siguientes parámetros.

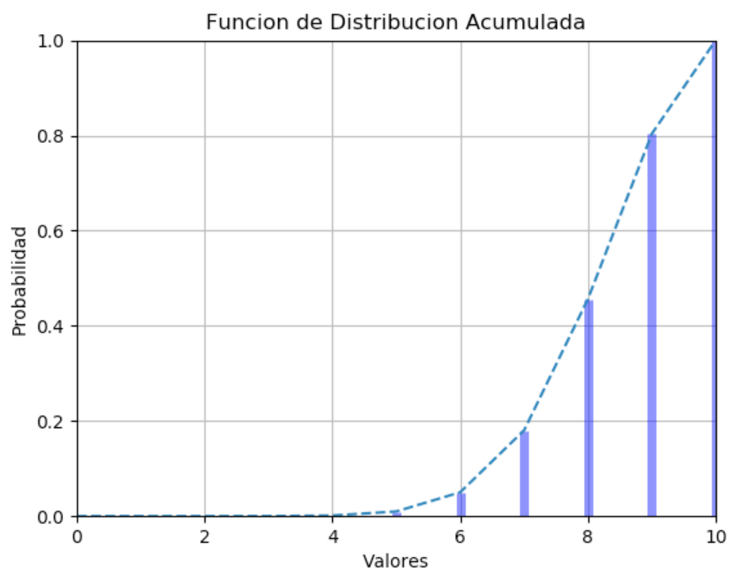
x = Cantidad de éxitos
n = Cantidad de ensayos = 10
p = Probabilidad de éxito = 8498 edades que no marcan error en una muestra de 10000 edades = 0.8498

La función de distribución respectiva corresponde a:

$$f(X = x) = \frac{10!}{x!\,(10! - x!)} \cdot (0.8498)^{x} \cdot (1 - 0.8498)^{10-x}$$

Mediante el ajuste de distribución se obtiene la siguiente grafica:





Resolviendo la problemática, mediante la distribución es posible obtener que la probabilidad que ninguna de error es de 0.196411.

<u>dataSet02</u>

"Este set de datos contienen 10.000 edades que se quieren cargar a una base de datos, sin embargo, la base da datos contiene como parámetro un NUMERIC de 2 dígitos, por lo tanto, todos aquellos valores de 3 dígitos marcaran error al momento de insertarlos en la base de datos. Calcule la probabilidad que de 10 edades seleccionadas, todas se puedan ingresar a la base de datos"

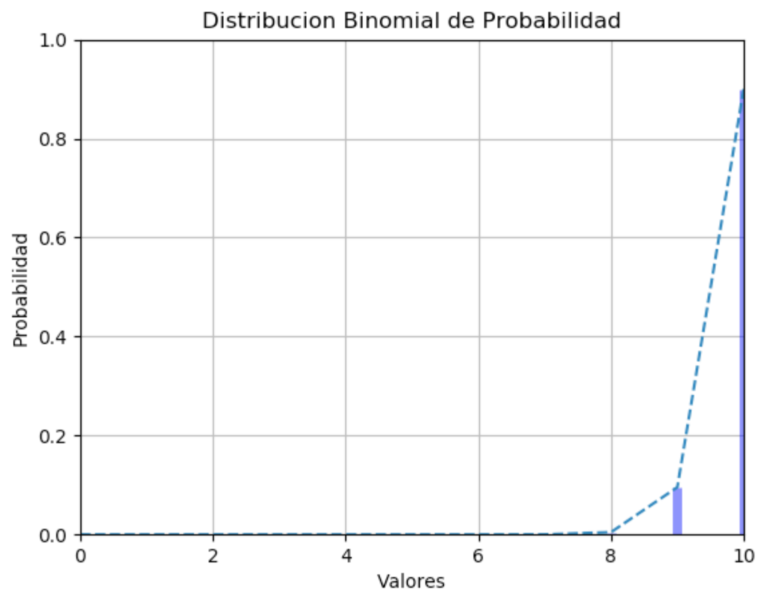
Al tratarse de una variable discreta, es posible plantear el set de datos mediante una distribución binomial, en donde se considera como éxito que la edad seleccionada no marque error (no contenga 3 dígitos), y como fallo que marque error (contenga 3 dígitos). Se obtendría los siguientes parámetros.

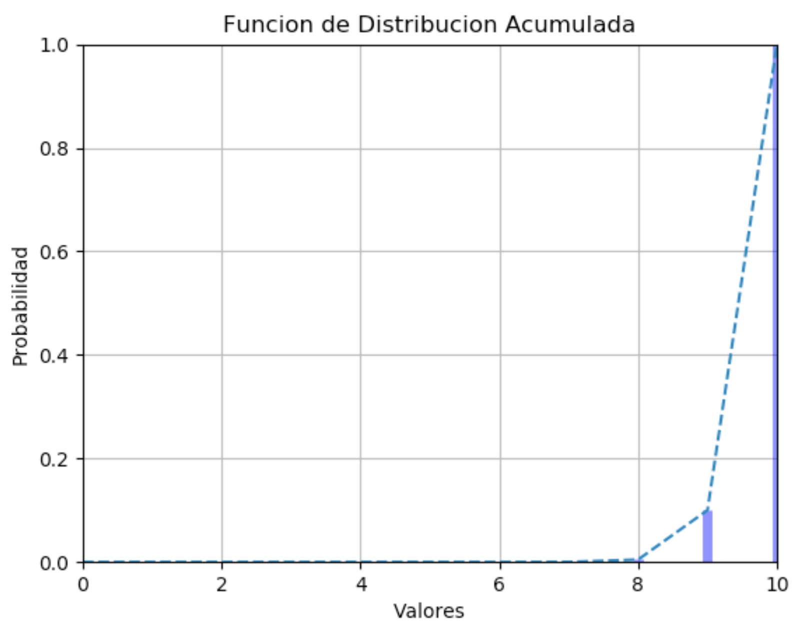
x = Cantidad de éxitos
n = Cantidad de ensayos = 10
p = Probabilidad de éxito = 9895 edades que no marcan error en una muestra de 10000 edades = 0.9895

La función de distribución respectiva corresponde a:

$$f(X = x) = \frac{10!}{x!\,(10! - x!)} \cdot (0.9895)^{x} \cdot (1 - 0.9895)^{10-x}$$

Mediante el ajuste de distribución se obtiene la siguiente grafica:





Mediante análisis es posible observar que al ser una probabilidad muy grande el extraer un dato que no contenga 3 dígitos, al tener una muestra de 10 datos, es muy poco probable que uno de los datos extraídos marque error, siendo lo mas probable que de los 10 datos extraídos todos se puedan ingresar a la base de datos, con una probabilidad de 0.899824.

dataSet03

"Este set de datos contiene 10000 tiempos de ejecución en segundos que tomo insertar un archivo de gran volumen a una base de datos con una tabla no indexada, ejecutado en equipos con un disco SSD. Calcule la probabilidad que de 40 equipos, mas de 10 tomaran un tiempo menor a los 10 min."

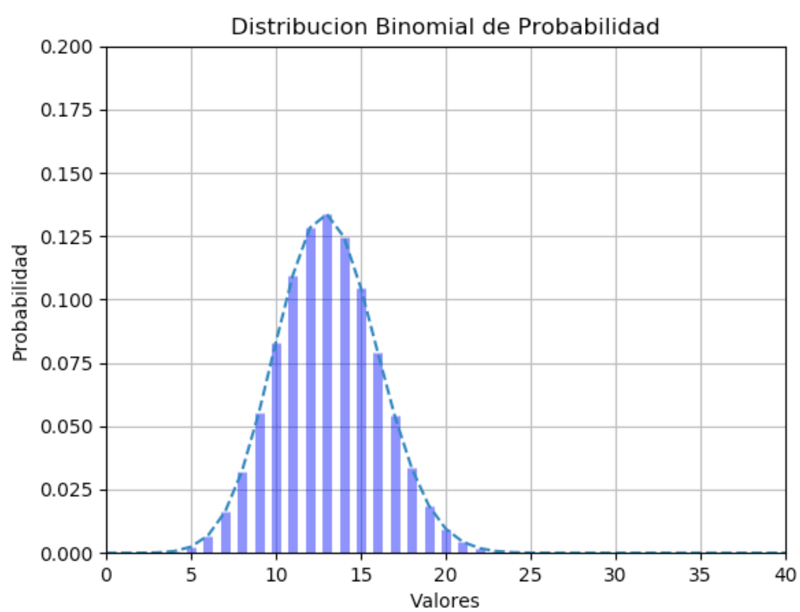
En primer lugar, es posible tomar del total de muestras, los datos que son menores a 10 min, obteniendo la probabilidad de éxito en una distribución binomial, cuyos parámetros son los siguientes:

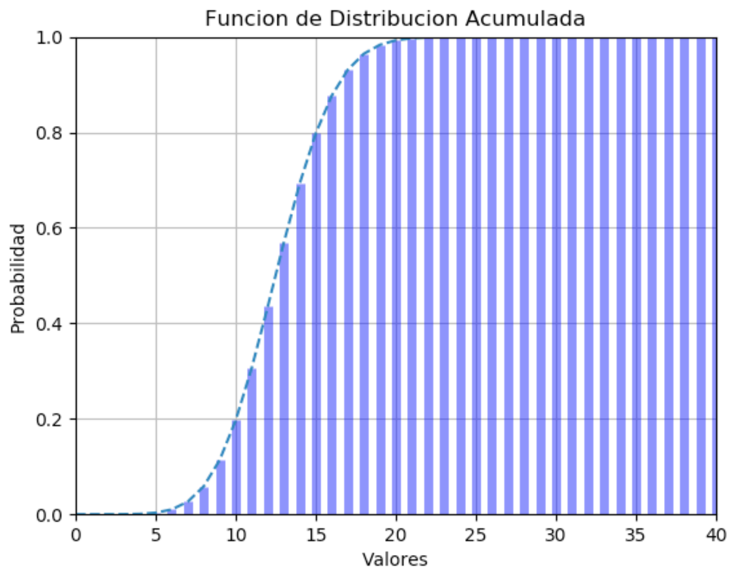
x = Cantidad de éxitos
n = Cantidad de ensayos = 40
p = Probabilidad de éxito = 3261 datos de un tiempo de ejecución menor a los 10 min en una muestra de 10000 datos = 0.3261

La función de distribución respectiva corresponde a:

$$f(X = x) = \frac{40!}{x!\,(40! - x!)} \cdot (0.3261)^{x} \cdot (1 - 0.3261)^{40-x}$$

Mediante el ajuste de distribución se obtiene la siguiente grafica:





En la distribución acumulada es posible ver que desde 5 computadores la probabilidad se comienza acumular mas rápido, teniendo como resultado que de los 40 extraídos, la probabilidad de que mas de 10 tomaran un tiempo menor de 10 minutos es de 0.886101

dataSet04

"Este set de datos contiene 10000 tiempos de ejecución en segundos que tomo insertar un archivo de gran volumen a una base de datos con una tabla indexada, ejecutado en equipos con un disco SSD. Calcule la probabilidad que de 15 equipos seleccionados, a lo mas 5 de ellos tomen un tiempo mayor a los 14 min."

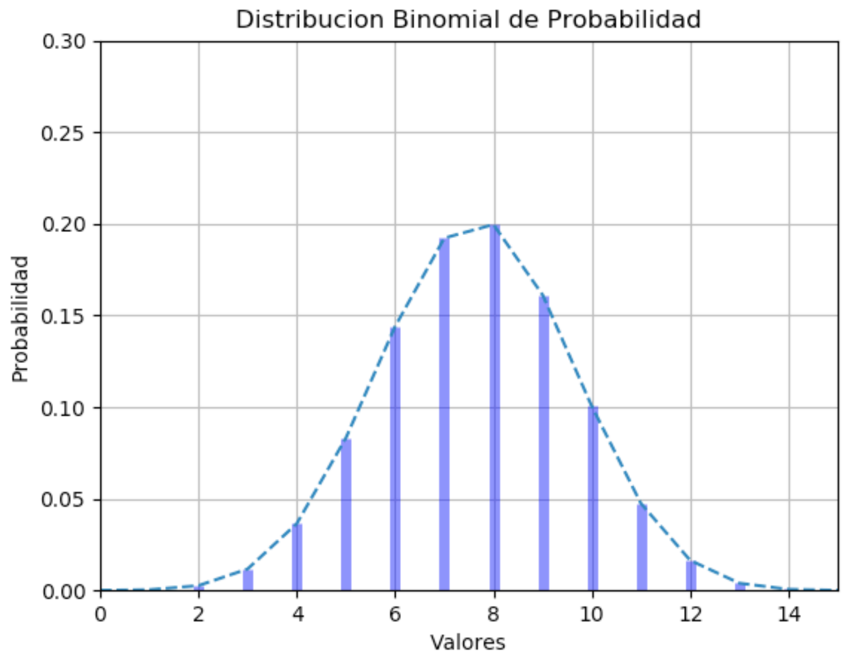
En primer lugar, es posible tomar del total de muestras, los datos que son mayores a 14 min, obteniendo la probabilidad de éxito en una distribución binomial, cuyos parámetros son los siguientes:

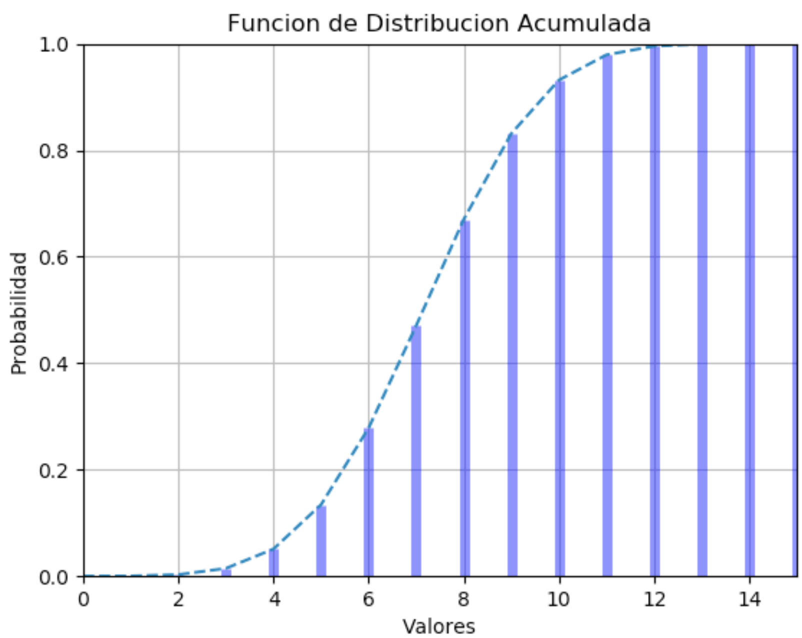
x = Cantidad de éxitos
n = Cantidad de ensayos = 15
p = Probabilidad de éxito = 5094 datos de un tiempo de ejecución mayor a los 14 min en una muestra de 10000 datos = 0.5094

La función de distribución respectiva corresponde a:

$$f(X = x) = \frac{15!}{x!\,(15! - x!)} \cdot (0.5094)^{x} \cdot (1 - 0.5094)^{15-x}$$

Mediante el ajuste de distribución se obtiene la siguiente grafica:





En la distribución acumulada es posible ver que desde la probabilidad acumulada hasta el valor 5 corresponde a la probabilidad de que a lo mas 5 tomaran un tiempo mayor de 14 minutos, siendo equivalente a 0.134295

dataSet05

"Este set de datos contiene 10000 tiempos de ejecución en segundos que tomo recuperar información de una base de datos con tabla no indexada, ejecutado en equipos con un disco SSD. Calcule la probabilidad que de 8 equipos, la mitad tome un tiempo mayor a los 50 seg."

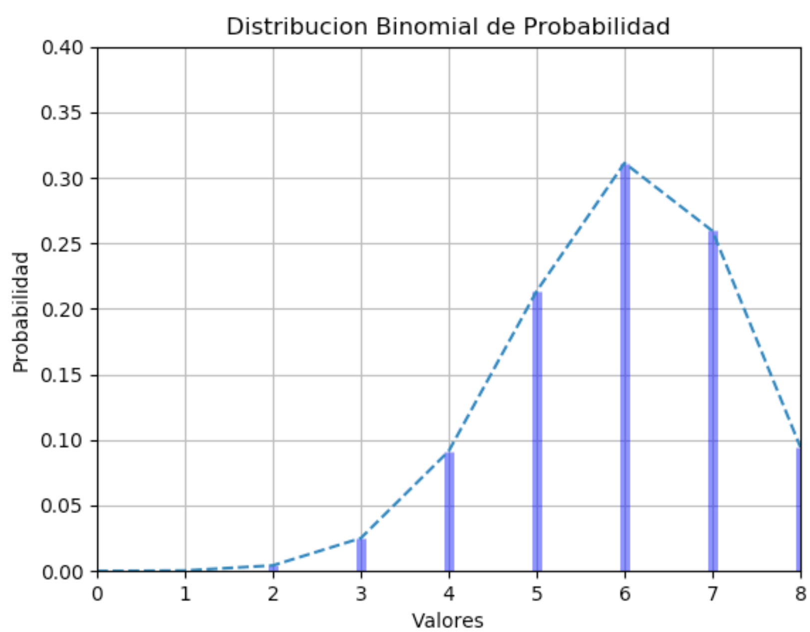
En primer lugar, es posible tomar del total de muestras, los datos que son mayores a 50 seg, obteniendo la probabilidad de éxito en una distribución binomial, cuyos parámetros son los siguientes:

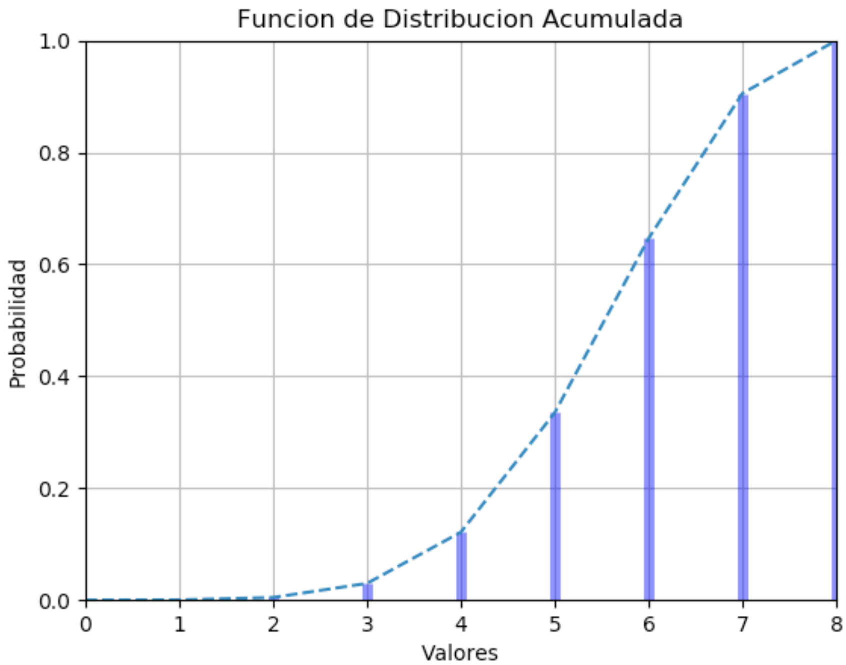
x = Cantidad de éxitos
n = Cantidad de ensayos = 8
p = Probabilidad de éxito = 7447 datos de un tiempo de ejecución mayor a los 50 seg en una muestra de 10000 datos = 0.7447

La función de distribución respectiva corresponde a:

$$f(X = x) = \frac{8!}{x!\,(8! - x!)} \cdot (0.7447)^{x} \cdot (1 - 0.7447)^{8-x}$$

Mediante el ajuste de distribución se obtiene la siguiente grafica:





Se obtiene como resultado que la probabilidad que la mitad de la muestra seleccionada correspondan a equipos que tuvieron un tiempo de ejecución mayor a los 50 segundos es de 0.091459

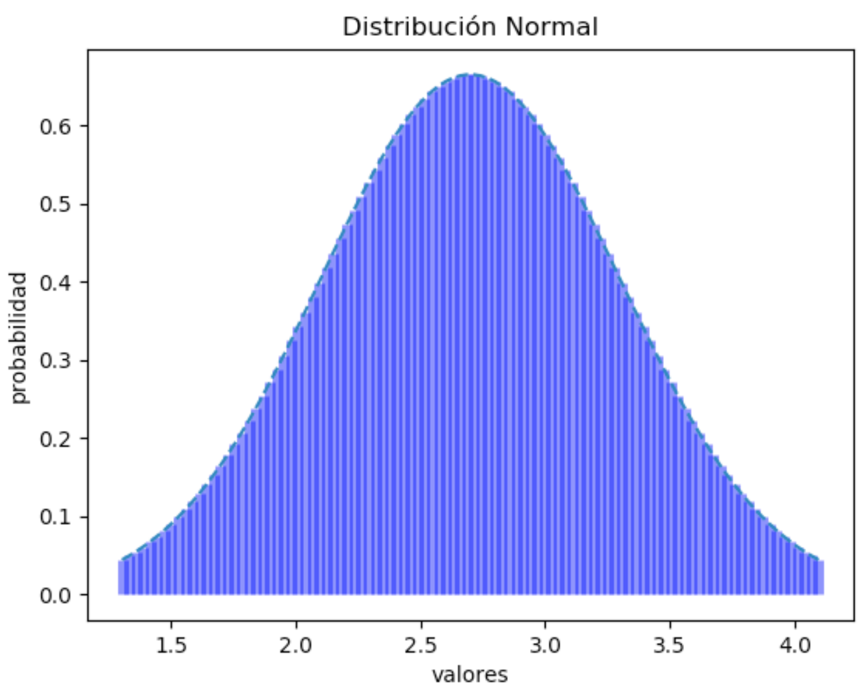
dataSet06

"Este set de datos contiene 10000 tiempos de ejecución en segundos que tomo recuperar información de una base de datos con tabla indexada, ejecutado en equipos con un disco SSD. Considerando que los datos no se diferencian mucho dentro de un rango, determine el porcentaje de equipos que tendrá un tiempo de ejecución que exceda los 3 segundos."

Al ser un set de datos con variables aleatorias continuas, es posible aplicar una distribución normal bajo sus parámetros de media y desviación estándar, los cuales son:

Media = 2.7
Desviación estándar = 0.6

Al aplicar el ajuste de distribución al set de datos, quedaría como resultado:



Ahora para obtener el porcentaje de equipos que tendrá un tiempo de ejecución que exceda los 3 segundos basta con obtener el valor correspondiente de Z en la distribución y determinar su probabilidad:

$$Z = \frac{x - \mu}{\sigma} = \frac{3 - 2.7}{0.2} = 0.5$$

$$P(X > 3) = P(Z > 0.5) = 1 - P(Z < 0.5) = 1 - 0.3085 = 0.6915$$

El porcentaje respectivo corresponde a 69.15%

dataSet07

"Este set de datos contiene 10000 tiempos de ejecución en segundos que tomo eliminar información de una base de datos con tabla no indexada, ejecutado en equipos con un disco SSD. Calcule la probabilidad que al tercer equipo seleccionado, tome un tiempo de ejecución mayor a 60 seg."

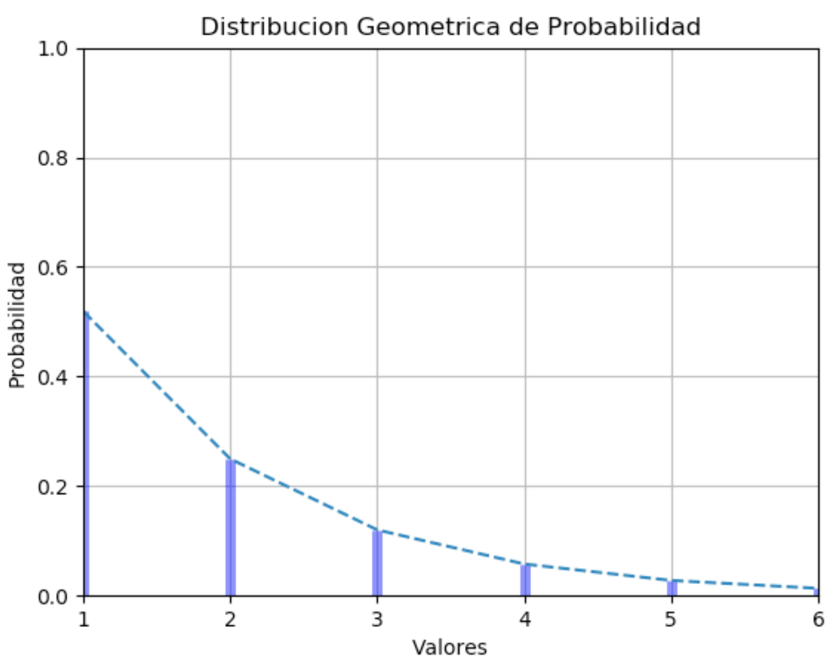
En primer lugar, es posible tomar del total de muestras, los datos que son mayores a 60 seg, obteniendo la probabilidad de éxito de un ensayo y establecer una distribución geométrica, cuyos parámetros son los siguientes:

x = Cantidad de ensayos necesarios
p = Probabilidad de éxito de un ensayo = 5190 datos de un tiempo de ejecución mayor a los 60 seg en una muestra de 10000 datos = 0.5190

La función de distribución correspondiente es:

$$f(X = x) = (1 - 0.5190)^{x} \cdot (0.5190)$$

Al aplicar el ajuste de distribución al set de datos, quedaría como resultado:



Como se puede observar, la probabilidad va disminuyendo cuando se aumentan los ensayos para obtener un tiempo de ejecución mayor a 60 segundos, en donde, se tiene como resultado que la probabilidad de obtener este datos al tercer equipo seleccionado es de 0.120076

dataSet08

"Este set de datos contiene 10000 tiempos de ejecución en segundos que tomo eliminar información de una base de datos con tabla no indexada, ejecutado con un disco SSD. Calcule la probabilidad que al quinto equipo seleccionado, tome un tiempo de ejecución menor a 10 seg."

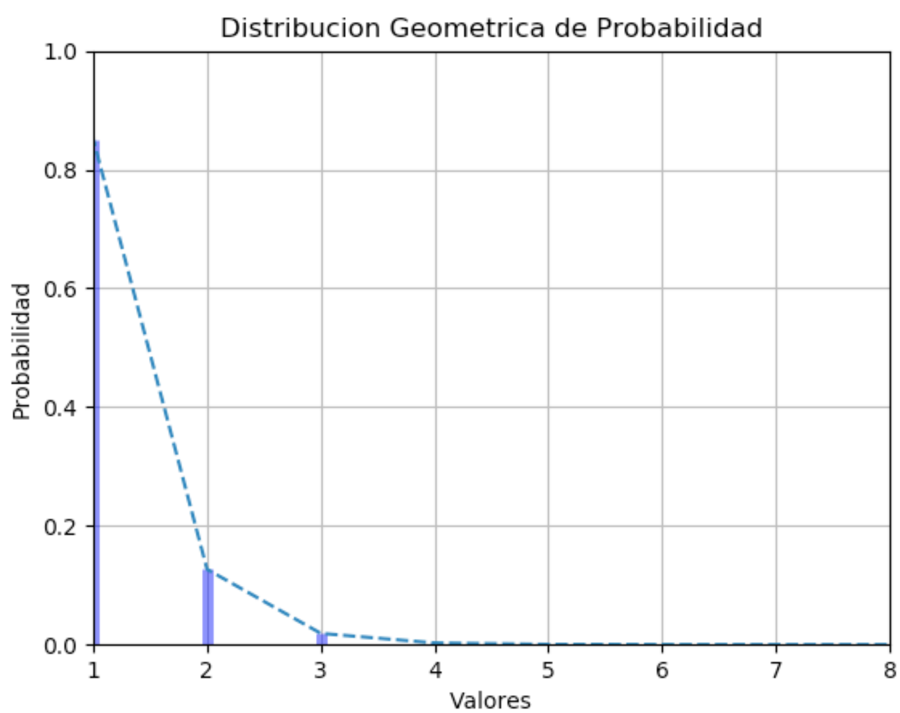
En primer lugar, es posible tomar del total de muestras, los datos que son menores a 10 seg, obteniendo la probabilidad de éxito de un ensayo y establecer una distribución geométrica, cuyos parámetros son los siguientes:

x = Cantidad de ensayos necesarios
p = Probabilidad de éxito de un ensayo = 8510 datos de un tiempo de ejecución menor a los 10 seg en una muestra de 10000 datos = 0.8510

La función de distribución correspondiente es:

$$f(X = x) = (1 - 0.8510)^{x} \cdot (0.8510)$$

Al aplicar el ajuste de distribución al set de datos, quedaría como resultado:



Como se puede observar, la probabilidad va disminuyendo cuando se aumentan los ensayos para obtener un tiempo de ejecución menor a 10 segundos, en donde, se tiene como resultado que la probabilidad de obtener este datos al quinto equipo seleccionado es de 0.000419.

dataSet09

"Este set de datos contiene 4000 tiempos de ejecución de un algoritmo de Fibonacci a valores entre 30 y 40 realizado mediante programación dinámica. Calcule la probabilidad que de 8 tiempos, la mitad tomaran mas de 0.0000015 seg."

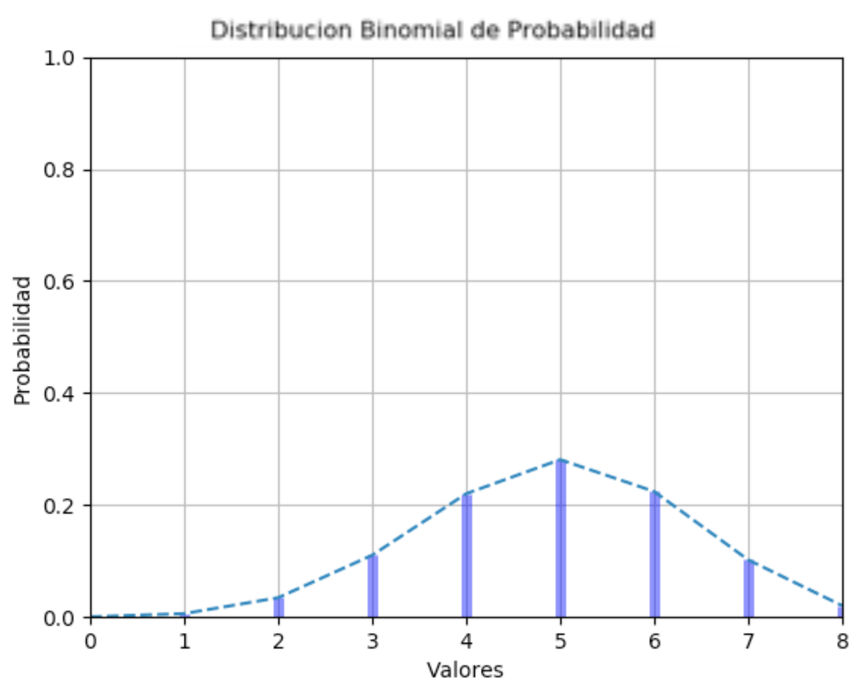
En primer lugar, es posible tomar del total de muestras, los datos que son mayores a 0.0000015 seg, obteniendo la probabilidad de éxito en una distribución binomial, cuyos parámetros son los siguientes:

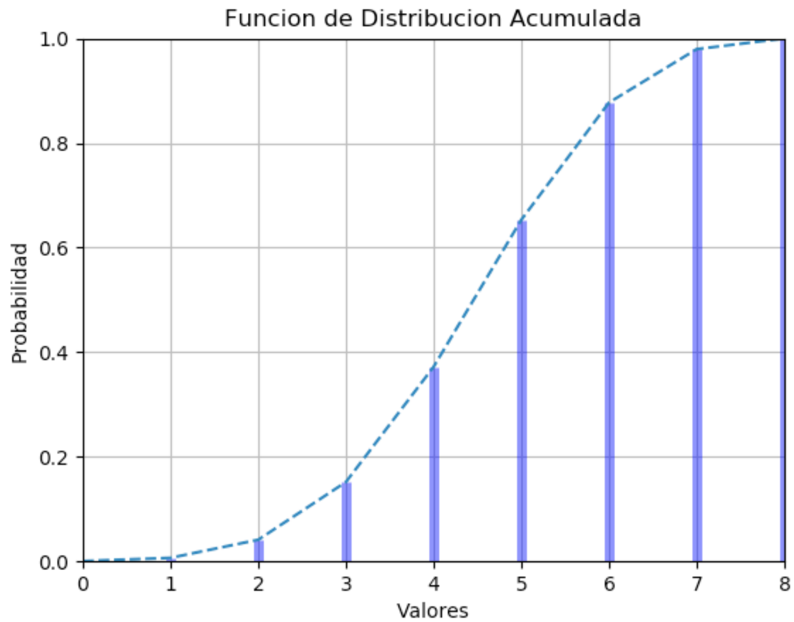
x = Cantidad de éxitos
n = Cantidad de ensayos = 8
p = Probabilidad de éxito = 2459 datos de un tiempo de ejecución mayor a los 0.0000015 seg en una muestra de 4000 datos = 0.6147

La función de distribución respectiva corresponde a:

$$f(X = x) = \frac{8!}{x!\,(8! - x!)} \cdot (0.6147)^{x} \cdot (1 - 0.6147)^{8-x}$$

Mediante el ajuste de distribución se obtiene la siguiente grafica:





Se obtiene como resultado que la probabilidad de que la mitad de la muestra seleccionada corresponda a tiempo de ejecución mayor a los 0.0000015 segundos es de 0.220265

dataSet10

“Este set de datos contiene 4000 tiempos de ejecución de un algoritmo de Fibonacci a valores entre 30 y 40 realizado mediante recursividad. Calcule la probabilidad que de 10 tiempos, la mitad tomaran mas de 10 seg.”

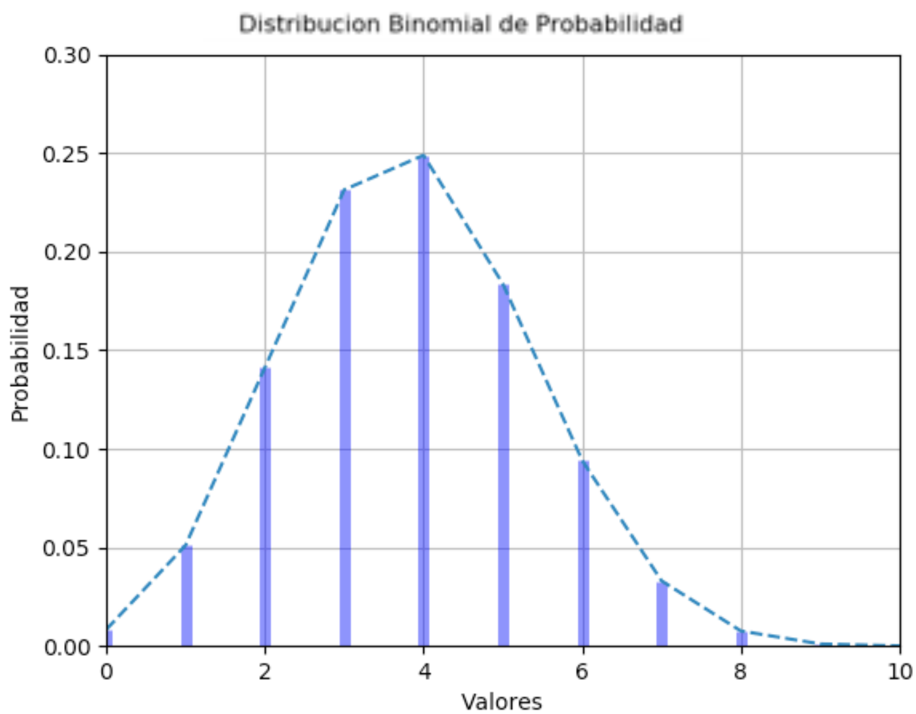
En primer lugar, es posible tomar del total de muestras, los datos que son mayores a 10 seg, obteniendo la probabilidad de éxito en una distribución binomial, cuyos parámetros son los siguientes:

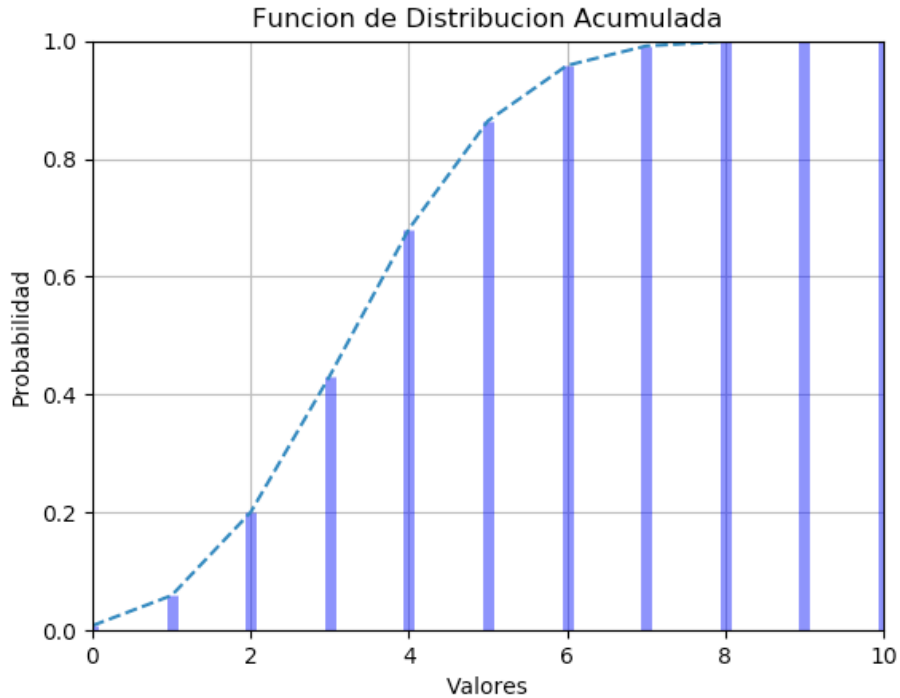
x = Cantidad de éxitos
n = Cantidad de ensayos = 10
p = Probabilidad de éxito = 1523 datos de un tiempo de ejecución mayor a los 10 seg en una muestra de 4000 datos = 0.3807

La función de distribución respectiva corresponde a:

$$f(X = x) = \frac{10!}{x!\,(10! - x!)} \cdot (0.3807)^{x} \cdot (1 - 0.3807)^{10-x}$$

Mediante el ajuste de distribución se obtiene la siguiente grafica:





Se obtiene como resultado que la probabilidad de que la mitad de la muestra seleccionada corresponda a un tiempo de ejecución mayor a los 10 segundos es de 0.183577

dataSet11

"Este set de datos contienes 4000 tiempos de ejecución de un algoritmos de Fibonacci a valores entre 30 y 40 realizado tanto en programación dinámica como recursivamente, considerando que con método recursivo los tiempos rodean entre 0.5 y 80 segundos, y con programación dinámica los tiempos no pasaron 0.1 segundos. Calcule la probabilidad que de 12 tiempos seleccionados, 4 o mas hayan sido realizado mediante método recursivo."

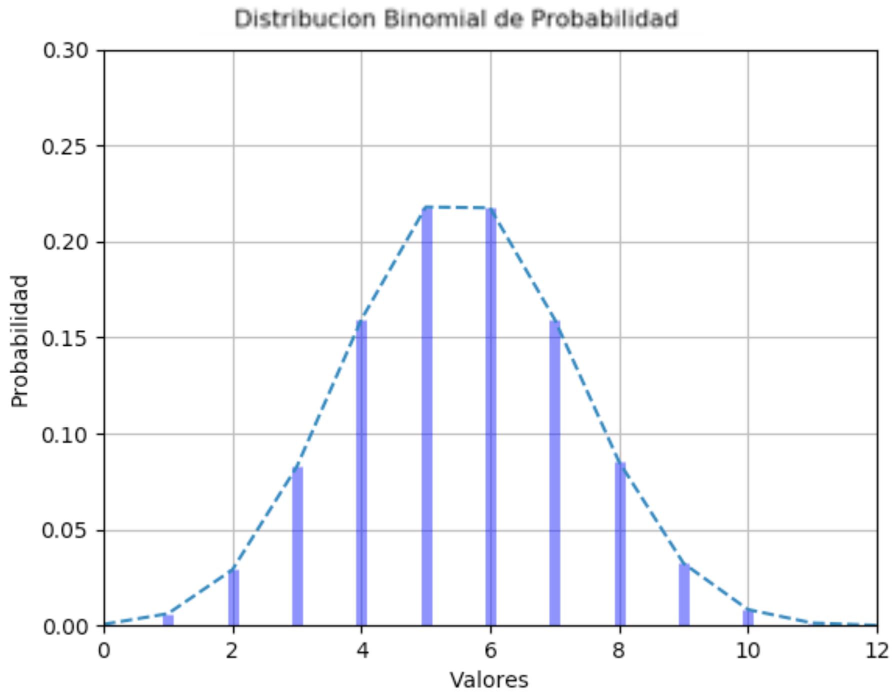
En primer lugar, es posible tomar del total de muestras, los datos que correspondan a tiempo realizados mediante método recursivo (mayores a 0.4 seg.) obteniendo la probabilidad de éxito en una distribución binomial, cuyos parámetros son los siguientes:

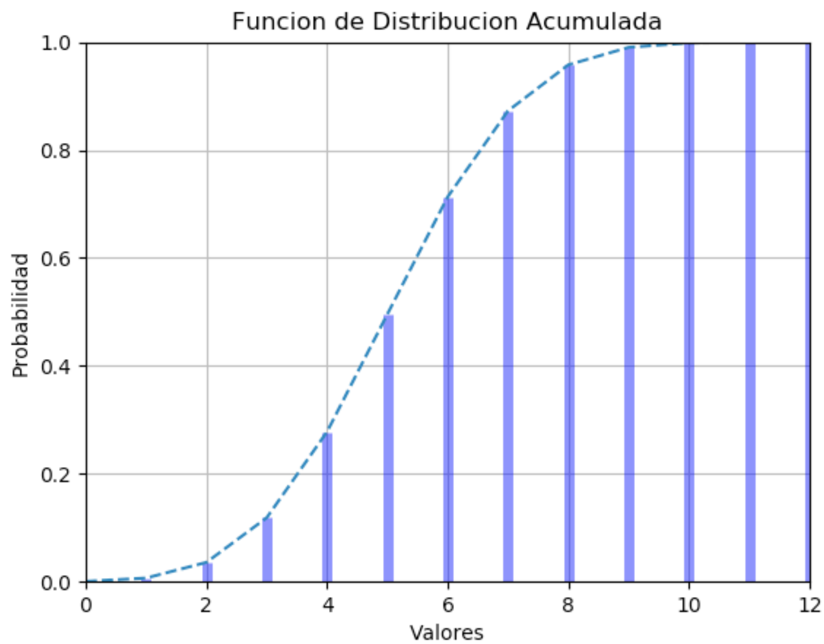
x = Cantidad de éxitos
n = Cantidad de ensayos = 12
p = Probabilidad de éxito = 1844 datos realizaron método recursivo en una muestra de 4000 datos = 0.461

La función de distribución respectiva corresponde a:

$$f(X = x) = \frac{12!}{x!\,(12! - x!)} \cdot (0.461)^{x} \cdot (1 - 0.461)^{12-x}$$

Mediante el ajuste de distribución se obtiene la siguiente grafica:





Como respuesta a la problemática, en la distribución acumulada es posible ver que la probabilidad que 4 o mas de los 12 extraídos hayan sido realizado mediante método recursivo es de 0.88143

dataSet12

"Este set de datos contiene 1000 tiempos de ejecución del algoritmo de ordenamiento Bubblesort, aplicado a listas de tamaños entre 10000 y 20000 valores random. Calcule la probabilidad que de 10 tiempos seleccionados, a los mas 6 tomen un tiempo menor a 30 seg."

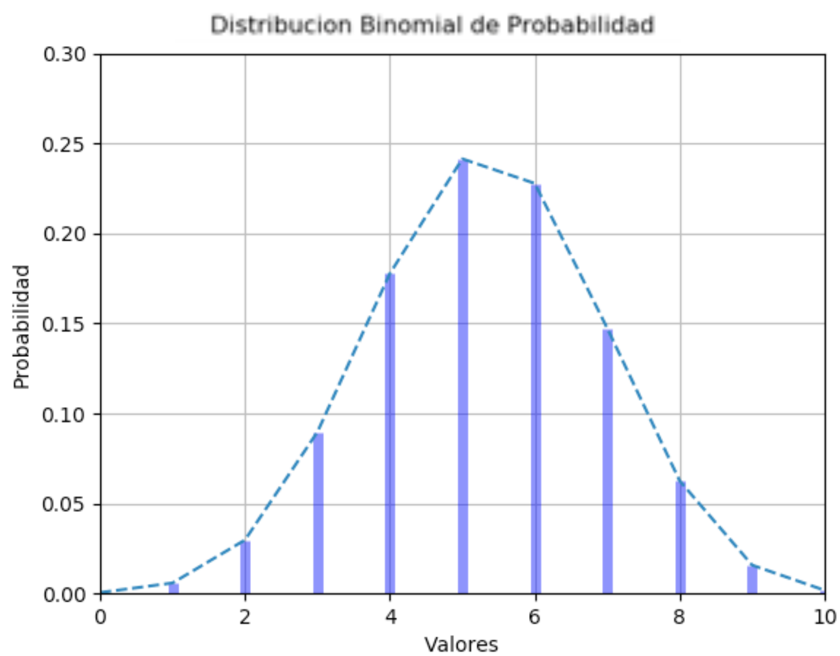
En primer lugar, es posible tomar del total de muestras, los datos que correspondan a un tiempo de ejecución menor a 30 seg, obteniendo la probabilidad de éxito en una distribución binomial, cuyos parámetros son los siguientes:

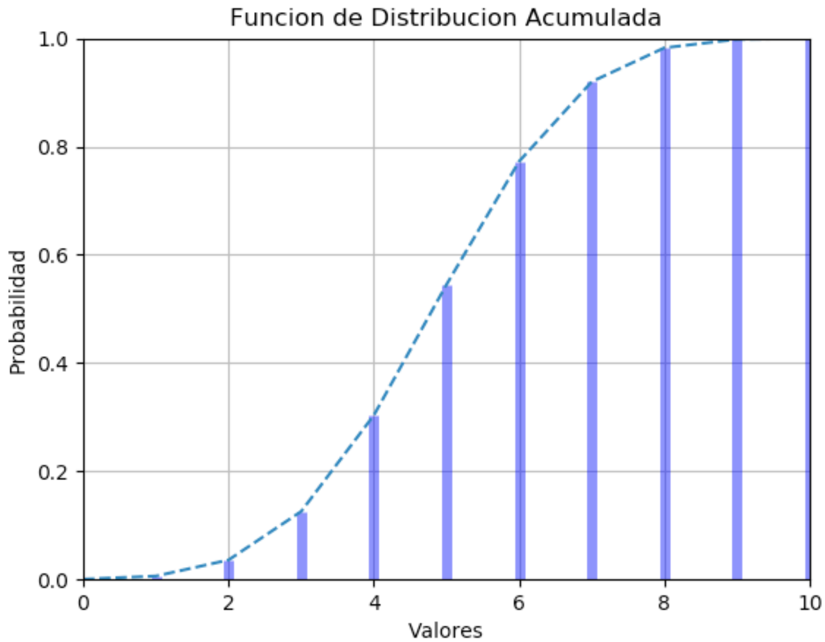
x = Cantidad de éxitos
n = Cantidad de ensayos = 10
p = Probabilidad de éxito = 531 datos con un tiempo menor a 30 seg. en una muestra de 1000 datos = 0.531

La función de distribución respectiva corresponde a:

$$f(X = x) = \frac{10!}{x!\,(10! - x!)} \cdot (0.531)^{x} \cdot (1 - 0.531)^{10-x}$$

Mediante el ajuste de distribución se obtiene la siguiente grafica:





Como respuesta a la problemática, en la distribución acumulada es posible ver que la probabilidad que a lo mas 6 datos de los 10 extraídos tomen un tiempo menor de 30 segundos es de 0.5442

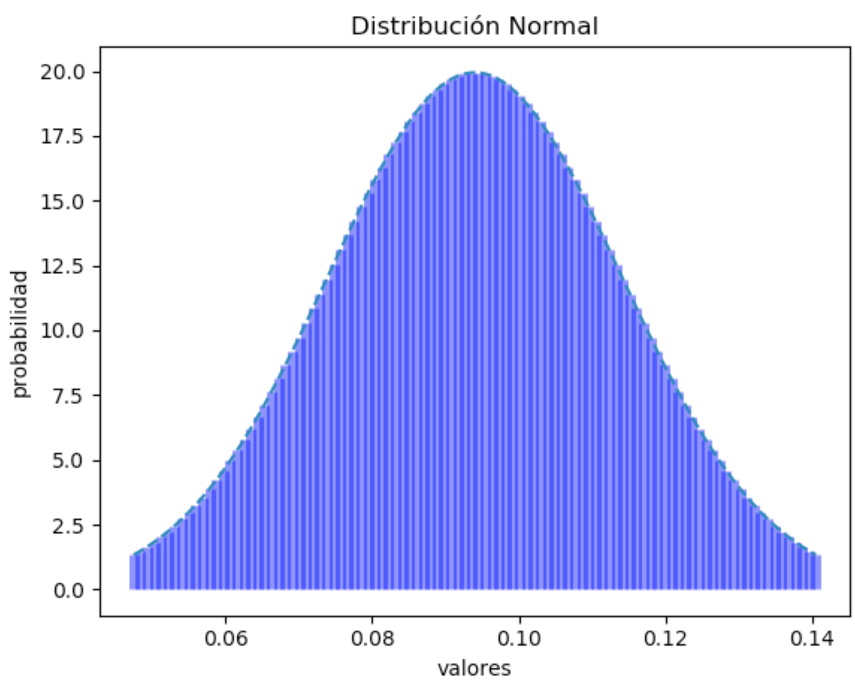
dataSet13

"Este set de datos contiene 1000 tiempos de ejecución del algoritmo de ordenamiento Mergesort, aplicado a listas de tamaños entre 10000 y 20000 valores random. Considerando que los datos no se diferencian mucho dentro de un rango, determine el porcentaje de datos que tendrá un tiempo de ejecución que exceda los 0.1 segundos."

Al ser un set de datos con variables aleatorias continuas, es posible aplicar una distribución normal bajo sus parámetros de media y desviación estándar, los cuales son:

Media = 0.094
Desviación estándar = 0.020

Al aplicar el ajuste de distribución al set de datos, quedaría como resultado:



Ahora para obtener el porcentaje de datos que tendrá un tiempo de ejecución que exceda los 0.1 segundos basta con obtener el valor correspondiente de Z en la distribución y determinar su probabilidad:

$$Z = \frac{x - \mu}{\sigma} = \frac{0.1 - 0.094}{0.020} = 0.3$$

$$P(X > 0.1) = P(Z > 0.3) = 1 - P(Z < 0.3) = 1 - 0.6179 = 0.38221$$

El porcentaje respectivo corresponde a 38.21%

dataSet14

“Este set de datos contiene 1000 tiempos de ejecución del algoritmo de ordenamiento Quicksort, aplicado a listas de tamaños entre 10000 y 20000 valores random. Calcule la probabilidad que recién el decimo tiempo seleccionado sea mayor a 0.1 seg.”

<u>dataSet15</u>

"Este set de datos contiene 1000 tiempos de ejecución tanto del algoritmo de ordenamiento Bubblesort como de Mergesort, aplicado a listas de tamaños entre 10000 y 20000 valores random, considerando que al arreglar las listas con Bubblesort siempre tomo mas de 10 segundos y con Mergesort no paso el segundo. Calcule la probabilidad de que seleccionando 8 tiempos, la mitad hayan sido realizados mediante el algoritmo Bubblesort."

dataSet16

"Este set contiene 284 valores de la variación de casos activos diario de covid en Chile (Datos extraídos de covid19.sporta.cl ACTUALIZACION DOMINGO 13 DE DICIEMBRE) Calcule la probabilidad de que al seleccionar 8 días, estos tuvieron una variación negativa."

<u>dataSet17</u>

“Este set contiene 284 valores de casos covid confirmados sintomáticos en un día en Chile (Datos extraídos de covid19.sporta.cl ACTUALIZACION DOMINGO 13 DE DICIEMBRE) Calcule la probabilidad que de 10 días seleccionados, existieron menos de 1000 casos.”

dataSet18

“Este set contiene 284 valores de casos recuperados de covid en un día en Chile (Datos extraídos de covid19.sporta.cl ACTUALIZACION DOMINGO 13 DE DICIEMBRE) Calcule la probabilidad que de 15 días seleccionados, existieron mas de 2000 casos.”

dataSet19

"Este set contiene 284 valores de tasa de hospitalización por covid en un día en Chile (Datos extraídos de covid19.sporta.cl ACTUALIZACION DOMINGO 13 DE DICIEMBRE). Calcule la probabilidad que de 8 días seleccionados, la tasa fue menor a 6."