DISTRIBUCIONES SET DE DATOS

(13 de dic. de 2020)

**dataSet01**

*“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.

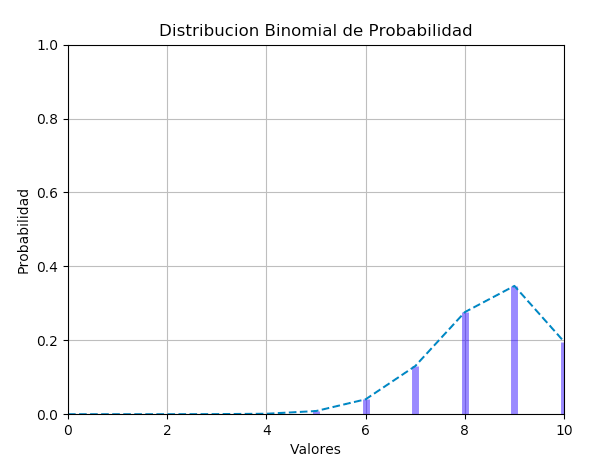
x = Cantidad de éxitos

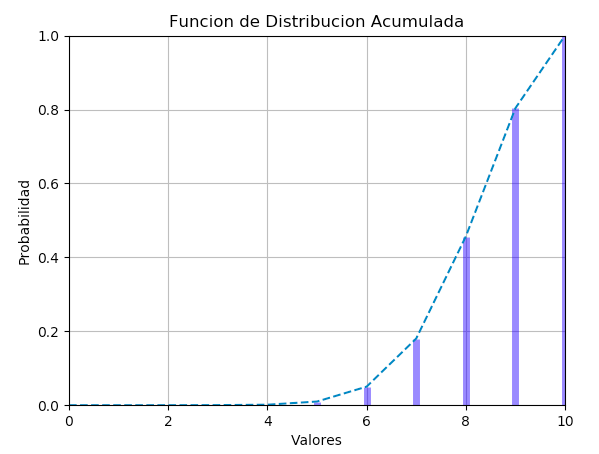
n = Cantidad de ensayos = 10

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:

Mediante el ajuste de distribución se obtiene la siguiente grafica:





Mediante análisis es posible observar que es muy poco probable que de la muestra de 10 edades, 4 o mas den error, y lo mas probable es que de las 10 extraídas, solo una de error, ahora resolviendo la problemática, mediante la distribución es posible obtener que la probabilidad que ninguna de error es de 0.196411.

**dataSet02**

*“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.

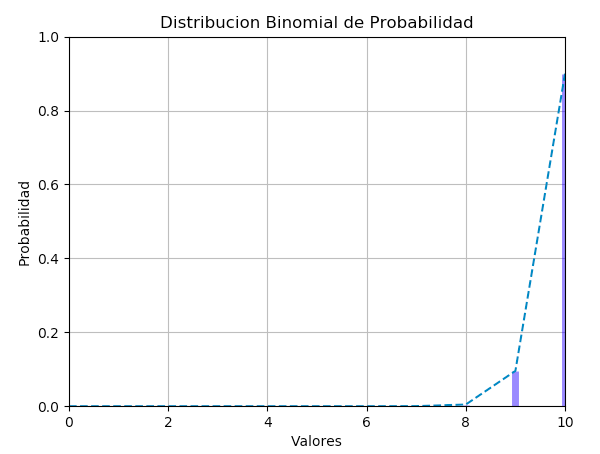
x = Cantidad de éxitos

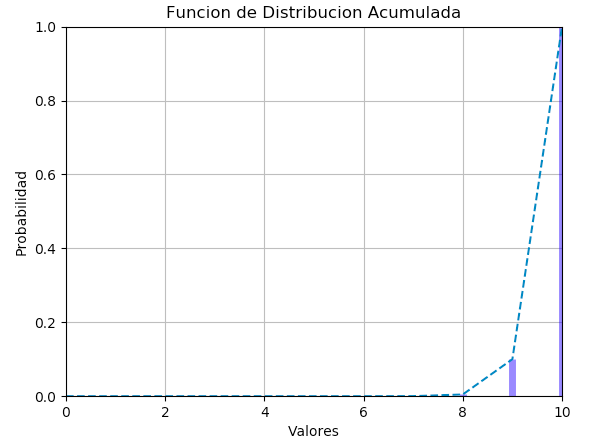
n = Cantidad de ensayos = 10

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:

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:

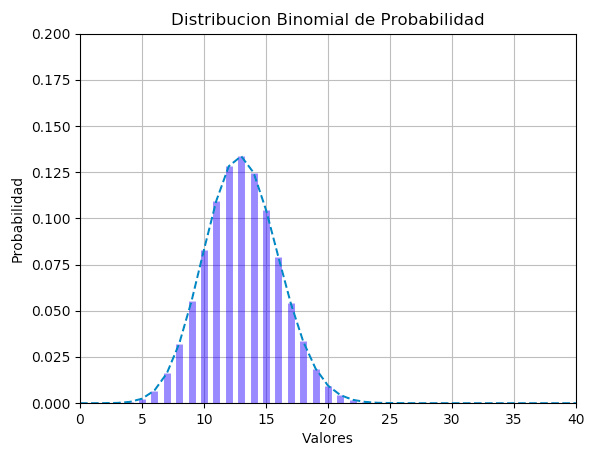
x = Cantidad de éxitos

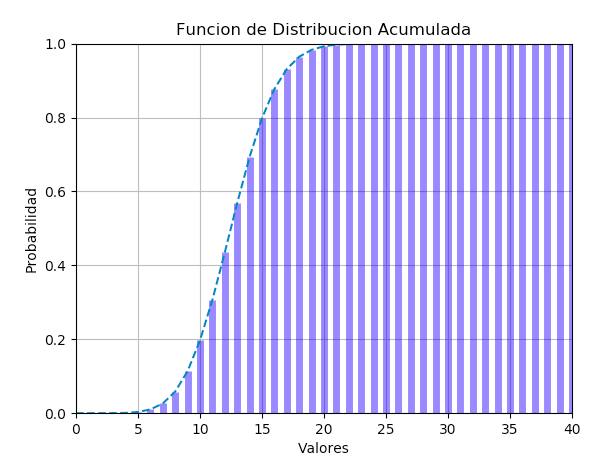
n = Cantidad de ensayos = 40

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:

Mediante el ajuste de distribución se obtiene la siguiente grafica:





Como se puede ver en la distribución, al extraer una muestra de 40 equipos, es mas probable que 13 tengan un tiempo de ejecución menor de 7 minutos en el contexto mencionado, por otro lado, 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:

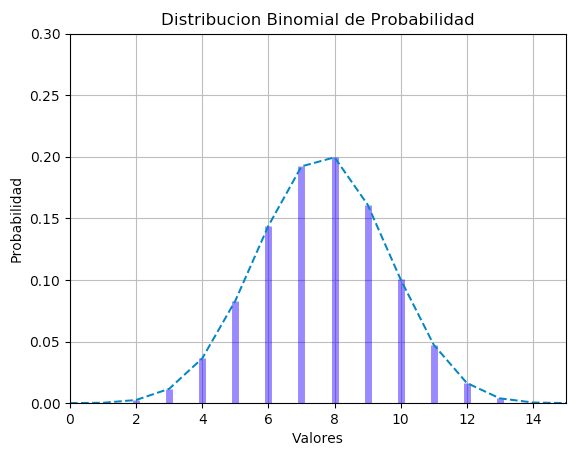
x = Cantidad de éxitos

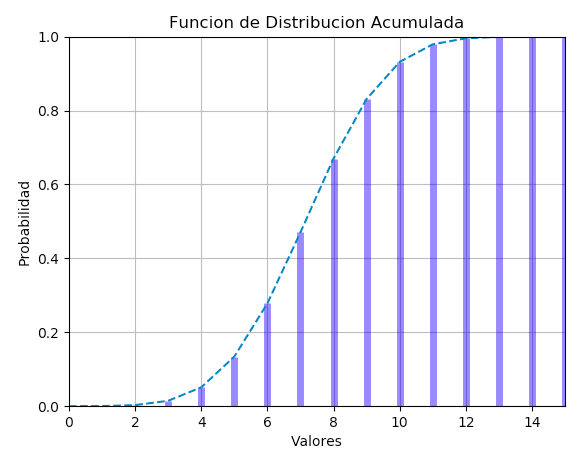
n = Cantidad de ensayos = 15

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:

Mediante el ajuste de distribución se obtiene la siguiente grafica:





Como se puede ver en la distribución, al extraer una muestra de 15 equipos, es mas probable que 8 tengan un tiempo de ejecución mayor de 14 minutos en el contexto mencionado, por otro lado, 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:

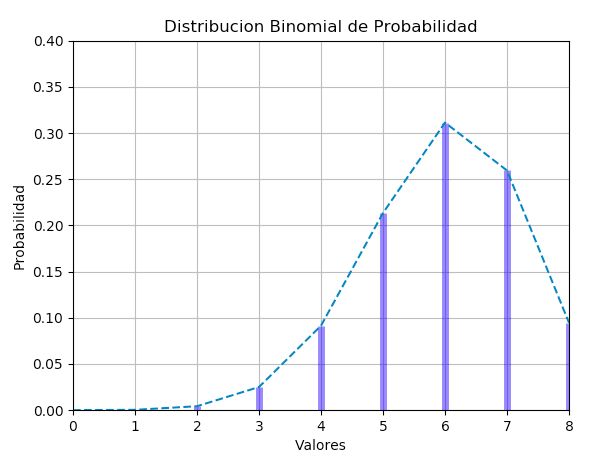
x = Cantidad de éxitos

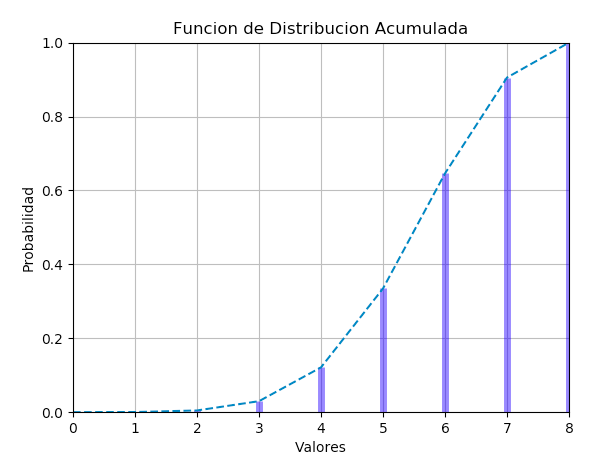
n = Cantidad de ensayos = 8

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:

Mediante el ajuste de distribución se obtiene la siguiente grafica:





Como se puede ver en la distribución, al extraer una muestra de 8 equipos, es mas probable que 6 tengan un tiempo de ejecución mayor de 50 segundos en el contexto mencionado, por otro lado, obtenemos como resultado que la probabilidad que la mitad de la muestra seleccionada correspondan a equipos que tuvieron un tiempo de ejecución menor a los 50 segundos es de 0.091459

**dataSet06**

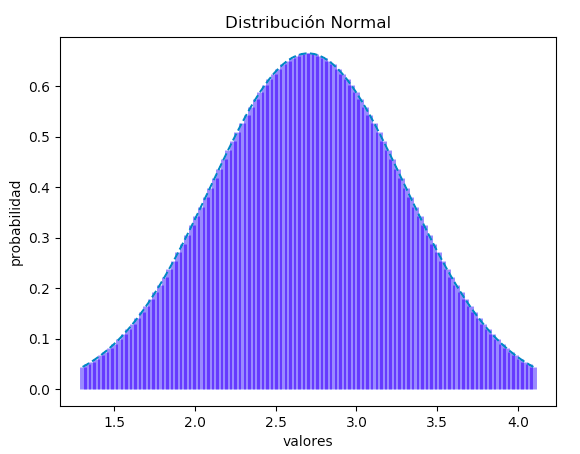
*“Este set de datos contiene 10000 tiempos de ejecución en segundos que tomo recuperar información de una base de datos con tabla indexada, ejecutado en equipos con un disco SSD. Considerando que los datos no se diferencian mucho dentro de un rango, determine el porcentaje de equipos que tendrá un tiempo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

El porcentaje respectivo corresponde a 30.85%

**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.”*

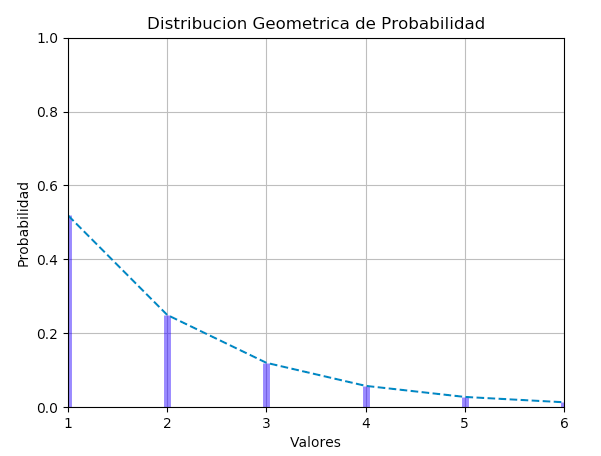
En primer lugar, es posible tomar del total de muestras, los datos que son mayores a 60 seg, obteniendo , la probabilidad de éxito de un ensayo y establecer una distribución geométrica, cuyos parámetros son los siguientes:

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:

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