Estadistica_24_Jun

January 31, 2018

0.0.1 Los datos son baratos pero el conocimiento es más difícil de conseguir

0.1 Estadística Descriptiva

Empezar a entender mis datos.

0.1.1 Media

Si se tiene una muestra de n valores: x_i La media μ es la suma de los valores dividido por el número de valores

$$\mu = \frac{1}{n}\sum_{i}^{n} x_i$$

```
In [1]: import pandas as pd
        import numpy as np
```

```
        data = pd.read_csv("train.csv")
        media_edad = np.mean(data['Age'])
```

```
        media_edad
```

```
Out[1]: 29.69911764705882
```

```
In [2]: data.describe()
```

```
Out[2]:        PassengerId    Survived      Pclass         Age       SibSp  \
        count   891.000000  891.000000  891.000000  714.000000  891.000000
        mean    446.000000    0.383838    2.308642   29.699118    0.523008
        std     257.353842    0.486592    0.836071   14.526497    1.102743
        min       1.000000    0.000000    1.000000    0.420000    0.000000
        25%     223.500000    0.000000    2.000000   20.125000    0.000000
        50%     446.000000    0.000000    3.000000   28.000000    0.000000
        75%     668.500000    1.000000    3.000000   38.000000    1.000000
        max     891.000000    1.000000    3.000000   80.000000    8.000000
```

```
              Parch        Fare
count   891.000000  891.000000
mean      0.381594   32.204208
std       0.806057   49.693429
min       0.000000    0.000000
25%       0.000000    7.910400
50%       0.000000   14.454200
75%       0.000000   31.000000
max       6.000000  512.329200
```

La media se encarga de describir la tendencia central de nuestros datos. ąImportante!, esta media μ se usa para describir a una población completa.

0.1.2 Varianza

Otro valor estadístico que nos ayuda a entender nuestros datos es la Varianza. A diferencia de la media que describe la tendencia de en donde se centran nuestros datos, la varianza describe que tan lejos se encuentran los datos de la media.

$$\sigma^2 = \frac{1}{n}\sum_{i}^{n}(x_i - \mu)^2$$

```
In [3]: varianza_edad = np.var(data['Age'])
        varianza_edad
```

```
Out[3]: 210.7235797536662
```

£Años al cuadrado? La varianza es difícil de interpretar debido a las unidades.
Por suerte la desviación estándar es un estadístico más significativo.

0.1.3 Desviación estándar

$$\sigma = \sqrt{\sigma}$$

```
In [4]: desviacion_edad = np.std(data['Age'])
        desviacion_edad
```

```
Out[4]: 14.516321150817317
```

```
In [6]: data.Age.std()
```

```
Out[6]: 14.526497332334044
```

ąImportante!, estas formulas para σ^2 y σ se usan para describir a una población completa.
Si lidiamos con una muestra de N valores se usan estimadores, $\bar{x}$ y S^2

$$\bar{x} = \frac{1}{N}\sum_{i}^{N} x_i$$

$$S^2 = \frac{1}{N-1}\sum_{i}^{N}(x_i - \bar{x})^2$$

0.2 Distribuciones

La media, la varianza y la desviación estándar son estadísticos concisos, pero también peligrosos, ya que nublan la información que nos proporcionan los datos.

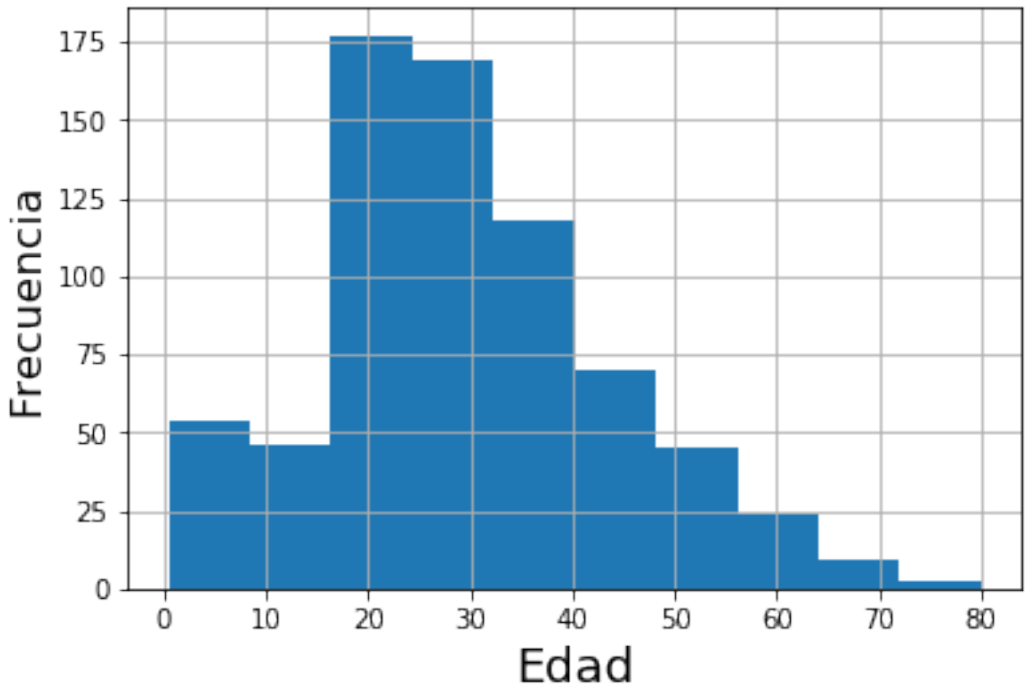
Un apoyo para entenderlos mejor es ver la distribución de los datos.

La representación más común de una distribución es un histograma, que describe frecuencia con la que aparece cada valor.

```
In [7]: %matplotlib inline
        import matplotlib.pyplot as plt

        edades = data[data.Age.notnull()]['Age']
        edades.hist()
        plt.xlabel('Edad', fontsize=18)
        plt.ylabel('Frecuencia', fontsize=16)
```

```
Out[7]: <matplotlib.text.Text at 0x11b489dd8>
```

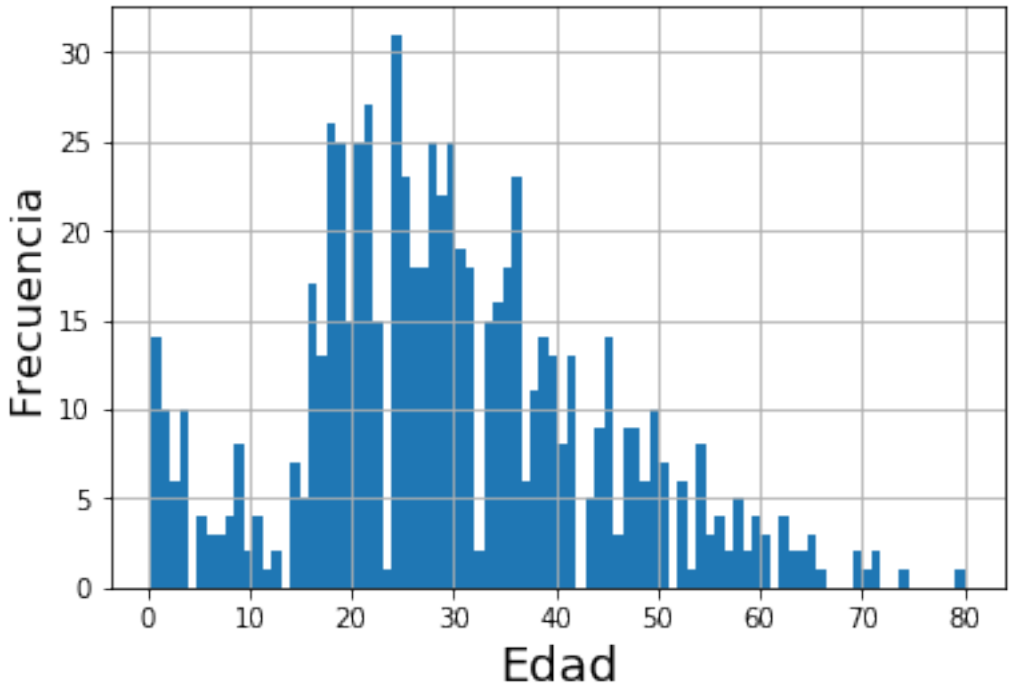



```
In [8]: len(edades.unique())
```

```
Out[8]: 88
```

```
In [9]: edades = data[data.Age.notnull()]['Age']
        edades.hist(bins=88)
        plt.xlabel('Edad', fontsize=18)
        plt.ylabel('Frecuencia', fontsize=16)
```

```
Out[9]: <matplotlib.text.Text at 0x11b634b70>
```



Los histogramas son útiles porque podemos revisar las siguientes características rápidamente:

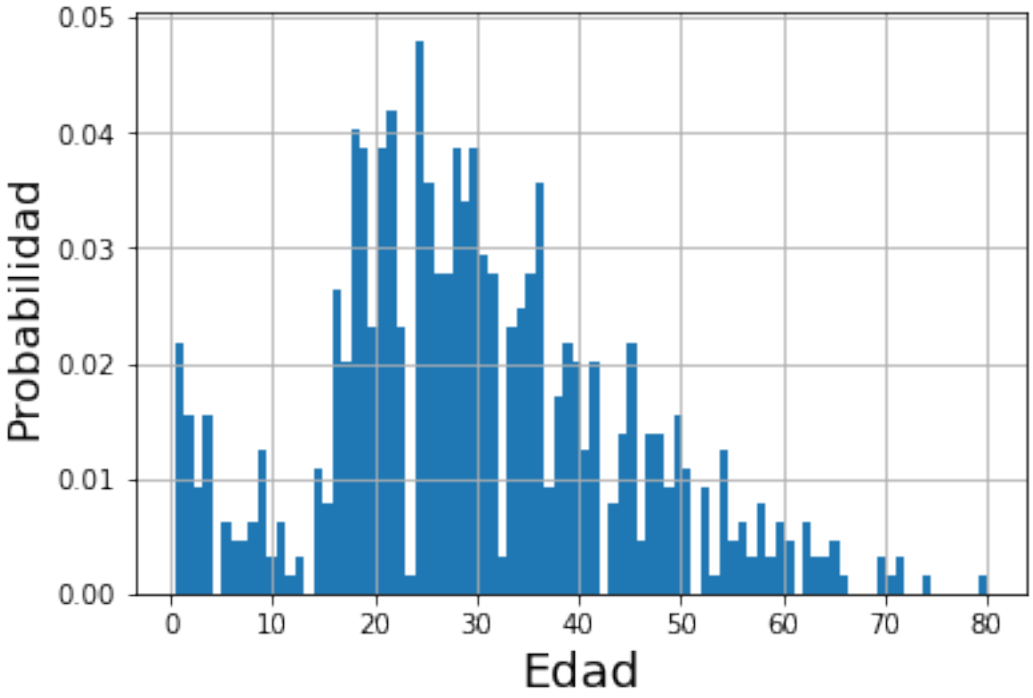
- Moda: El valor más común o que más se repite en una distribución se llama moda.
- Forma: Alrededor de la moda podemos ver que la distribución es asimétrica.
- Los valores atípicos. (outliers)

0.2.1 Función de probabilidad

Si queremos transformar las frecuencias a una función de probabilidad debemos dividir la serie entre el número de elementos

```
In [10]: edades.hist(bins=88, normed=True)
         plt.xlabel('Edad', fontsize=18)
         plt.ylabel('Probabilidad', fontsize=16)
```

```
Out[10]: <matplotlib.text.Text at 0x11b836f98>
```



$$P(X = x) = f(x)$$

La función de probabilidad funciona bien si el número de valores es pequeño.

Pero a medida que el número de valores aumenta, la probabilidad asociada a cada valor se hace más pequeño.

Una alternativa es usar la función de distribución acumulada.

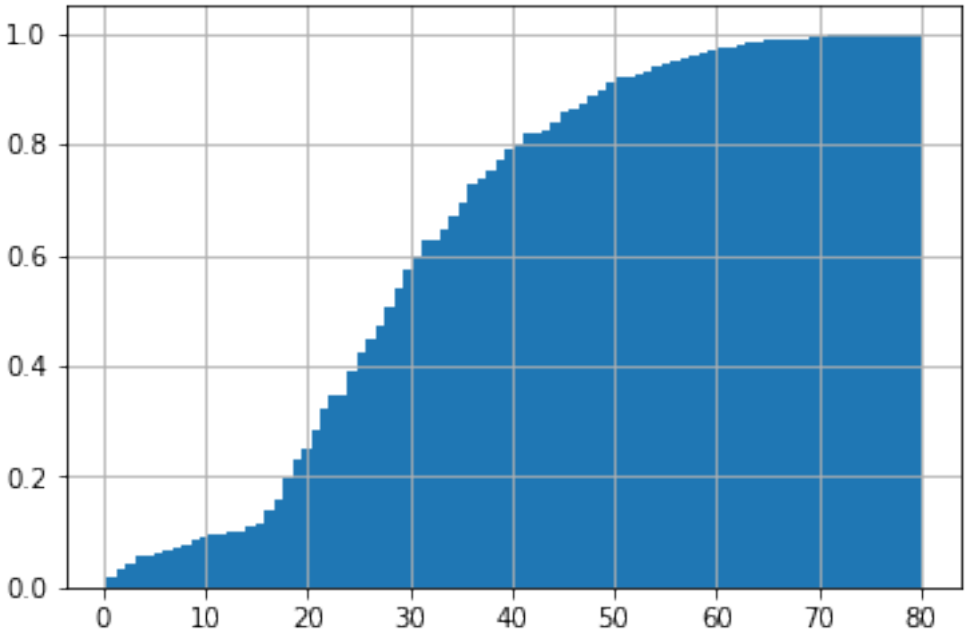
```
In [11]: edades.describe()
```

```
Out[11]: count    714.000000
         mean      29.699118
         std       14.526497
         min        0.420000
         25%       20.125000
         50%       28.000000
         75%       38.000000
         max       80.000000
         Name: Age, dtype: float64
```

0.2.2 Función de distribución acumulada

```
In [12]: edades.hist(cumulative=True, bins=88, normed=True)
```

```
Out[12]: <matplotlib.axes._subplots.AxesSubplot at 0x11bb3c198>
```



$$P(X \leq x) = f(x)$$

0.2.3 Distribuciones continuas

Cuando tenemos variables aleatorias continuas

La distribución normal (Gaussiana), es una de las distribuciones de probabilidad que con más frecuencia aparece aproximada en fenómenos reales.

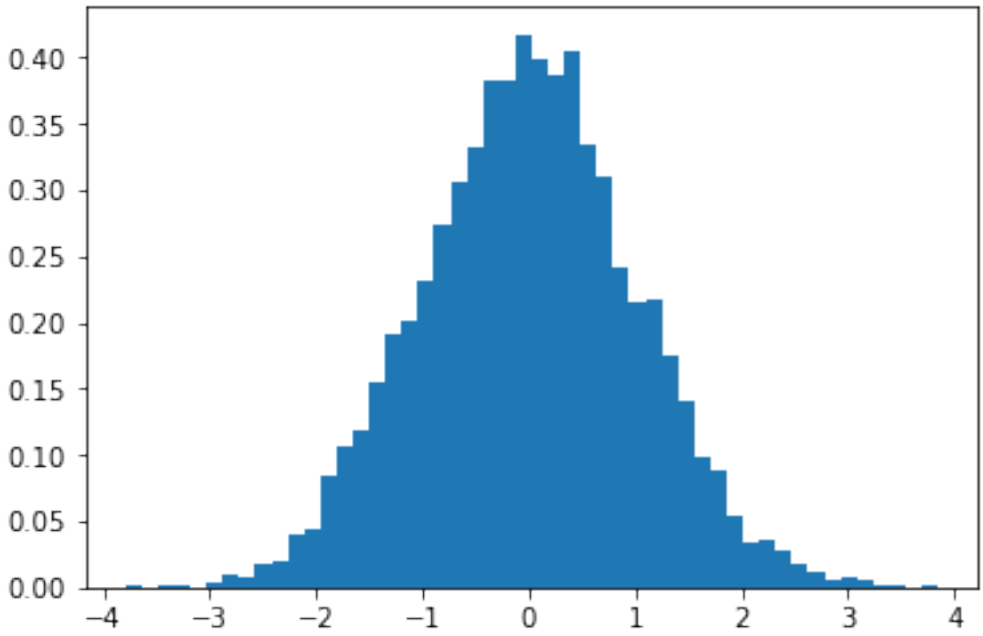
$$P(x, \sigma, \mu) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

```
In [13]: import numpy as np
         x = np.random.randn(5000)

         # Make a normed histogram. It'll be multiplied by 100 later.
         plt.hist(x, bins=50, normed=True)
```

```
Out[13]: (array([ 0.00130883,  0.        ,  0.00130883,  0.00130883,  0.        ,
                  0.00392649,  0.01047063,  0.00785297,  0.0183236 ,  0.01963243,
                  0.0405737 ,  0.04319135,  0.08376505,  0.10732397,  0.1177946 ,
                  0.15575064,  0.19108902,  0.20025082,  0.23166272,  0.27354524,
                  0.30626596,  0.33244254,  0.38217804,  0.38217804,  0.41751642,
                  0.39788399,  0.38741336,  0.40442813,  0.3350602 ,  0.31019245,
                  0.24082452,  0.21464794,  0.2172656 ,  0.17538307,  0.14004469,
                  0.09816217,  0.08900037,  0.05366199,  0.03402955,  0.03533838,
                  0.02748541,  0.0183236 ,  0.01177946,  0.00523532,  0.00785297,
```

```
        0.00523532,  0.00261766,  0.00130883,  0.        ,  0.00130883]),
array([-3.79367179, -3.64086343, -3.48805507, -3.33524671, -3.18243835,
       -3.02962998, -2.87682162, -2.72401326, -2.5712049 , -2.41839654,
       -2.26558818, -2.11277982, -1.95997146, -1.8071631 , -1.65435474,
       -1.50154637, -1.34873801, -1.19592965, -1.04312129, -0.89031293,
       -0.73750457, -0.58469621, -0.43188785, -0.27907949, -0.12627113,
        0.02653724,  0.1793456 ,  0.33215396,  0.48496232,  0.63777068,
        0.79057904,  0.9433874 ,  1.09619576,  1.24900412,  1.40181249,
        1.55462085,  1.70742921,  1.86023757,  2.01304593,  2.16585429,
        2.31866265,  2.47147101,  2.62427937,  2.77708773,  2.9298961 ,
        3.08270446,  3.23551282,  3.38832118,  3.54112954,  3.6939379 ,
        3.84674626]),
<a list of 50 Patch objects>)
```



0.2.4 £Por qué usar distribuciones continuas?

Como todos los modelos, las distribuciones continuas son abstracciones, lo que significa que pueden simplificar y deshacerse de los detalles que se consideran irrelevantes (Errores de medición, outliers).

Además son una forma de comprimir los datos. Ya que si logramos ajustar un modelo a un conjunto de datos, un pequeño conjunto de parámetros puede resumir una gran cantidad de datos.

0.2.5 £Por qué es tan importante la distribución Normal?

El teorema de límite central establece que la media de la muestra $\bar{X}$ sigue una distribución normal (para n grandes) con media μ y desviación estándar $\frac{\sigma}{\sqrt{(n)}}$

El teorema del límite central explica, porque aparece con tanta frecuencia la distribución normal en el mundo natural.

La mayoría de las características de los animales y otras formas de vida se ven afectadas por un gran número de variables genéticas y ambientales cuyo efecto es aditivo.

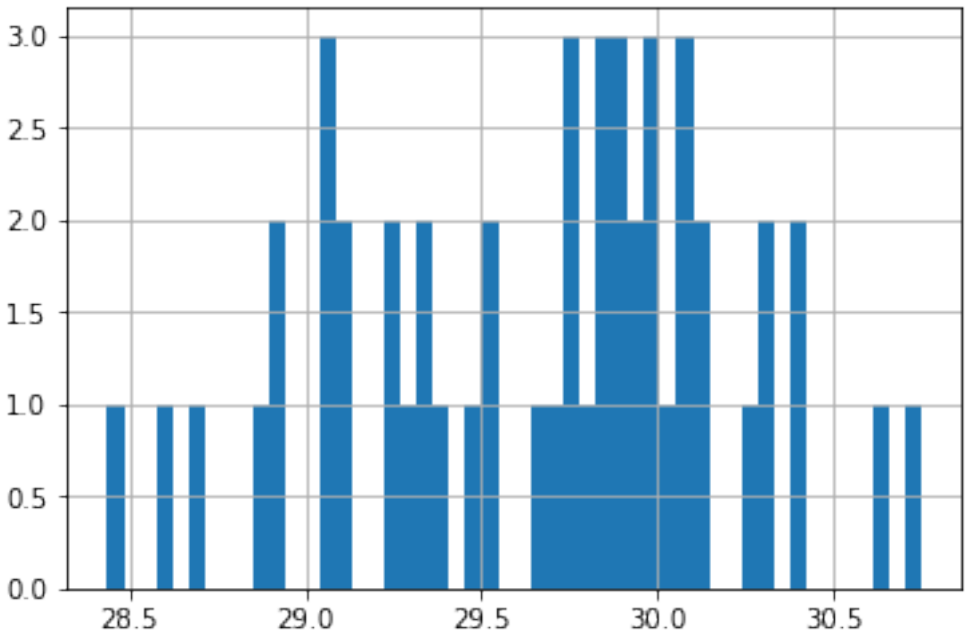
Las características que medimos son la suma de un gran número de pequeños efectos, por lo que su distribución tiende a ser normal.

```
In [14]: #Prueba para el Teorema de Limite Central usando 50
         media_muestra = [] #Iniciamos una lista

         for x in range(0, 50):
             media_muestra.append(np.mean(edades.sample(n=300)))

         media_muestra = pd.Series(media_muestra)
         media_muestra.hist(bins=50)
```

```
Out[14]: <matplotlib.axes._subplots.AxesSubplot at 0x11c067278>
```

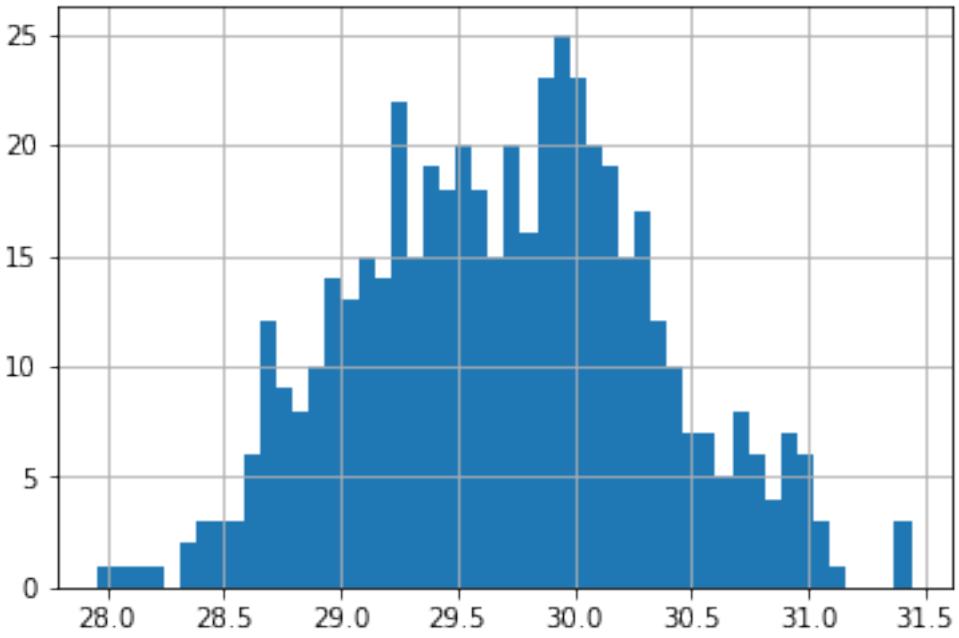



```
In [17]: #Prueba para el Teorema de Limite Central usando 500
         media_muestra = []

         for x in range(0, 500):
             media_muestra.append(np.mean(edades.sample(n=300)))

         media_muestra = pd.Series(media_muestra)
         media_muestra.hist(bins = 50)
```

Out[17]: <matplotlib.axes._subplots.AxesSubplot at 0x11c6d02b0>

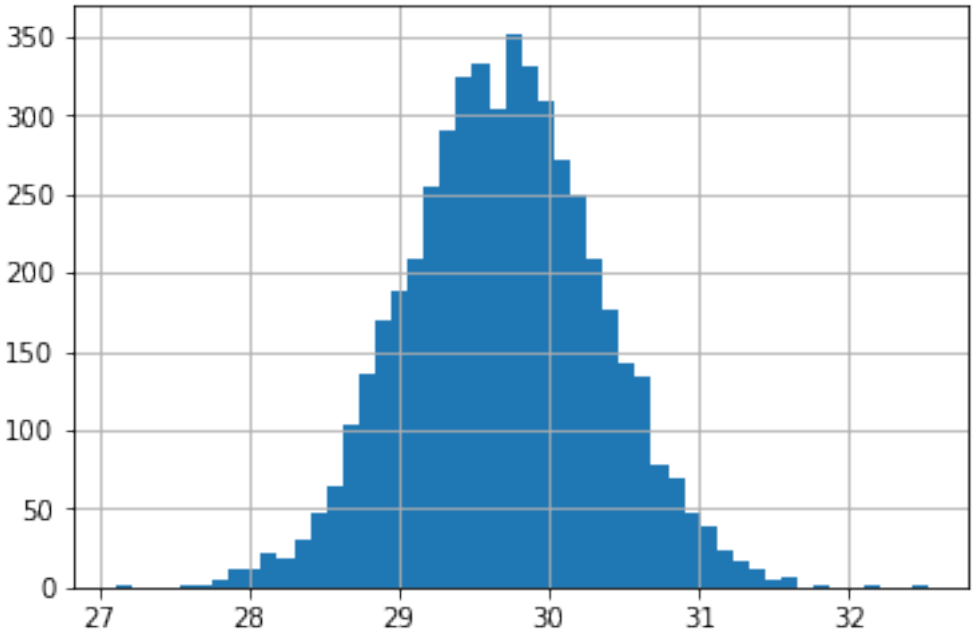


In [16]:
```
#Prueba para el Teorema de Limite Central usando 5000
media_muestra = []

for x in range(0, 5000):
    media_muestra.append(np.mean(edades.sample(n=300)))

media_muestra = pd.Series(media_muestra)
media_muestra.hist(bins = 50)
```

Out[16]: <matplotlib.axes._subplots.AxesSubplot at 0x11c41ef28>



0.2.6 Probabilidad

Anteriormente mencionamos que la probabilidad es la frecuencia expresada como una fracción tamaño de muestra.

Esa es una definición de probabilidad, pero no es la única y de hecho, el significado de probabilidad es un tema controversial.

Existe un consenso general de que la probabilidad es un valor real entre 0 y 1. Este valor pretende dar una medida cuantitativa que corresponde a la noción de que algunas cosas son más probables que otras.

$$P(E)\epsilon[0,1]$$

0.2.7 Reglas de probabilidad (Recordando a Kolmogorov)

- La probabilidad de que ocurra un evento es un valor entre 0 y 1. Para todo evento existe una probabilidad
$$0 \leq P(E) \leq 1$$
- La probabilidad de que nada ocurra es 0
$$P(\emptyset) = 0$$
- La probabilidad de que algo ocurra es 1
$$P(\Omega) = 1$$
- La probabilidad de algo es 1 menos la probabilidad de lo contrario

0.2.8 Probabilidad condicional

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Si A y B son eventos independientes entonces:

$$P(A|B) = \frac{P(A)P(B)}{P(B)} = P(A)$$

Monty Hall

```
In [ ]: from IPython.display import HTML
```

```
In [ ]: HTML('<iframe width="560" height="315" src="https://www.youtube.com/embed/mhlc7peGlGg?
```

0.2.9 Regla de Bayes

El teorema de Bayes es a menudo interpretado como una declaración acerca de cómo la evidencia, E, afecta la probabilidad de una hipótesis, H:

$$P(H|E) = P(H)\frac{P(E|H)}{P(E)}$$

En palabras, esta ecuación dice que la probabilidad de H después de haber visto E es el producto de $P(H)$, que es la probabilidad de que H antes de ver la evidencia E, y la relación de $P(E|H)$, la probabilidad de ver la evidencia suponiendo que H es verdadera, y $P(E)$, la probabilidad de ver la evidencia bajo cualquier circunstancia.

Ejemplo: Filtro de Spam

$$P(S|W) = \frac{P(W|S) \cdot P(S)}{P(W|S) \cdot P(S) + P(W|H) \cdot P(H)}$$

donde:

- $P(S|W)$ Es la probabilidad de que nuestro mensaje sea SPAM, sabiendo que encontramos la palabra "Dinero"
- $P(S)$ Es la probabilidad de que cualquier mensaje sea SPAM
- $P(W|S)$ La probabilidad de que nuestra palabra aparezca en mensajes de SPAM
- $P(H)$ La probabilidad de que nuestro mensaje sea HAM
- $P(W|H)$ La probabilidad de que la nuestra palabra aparezca en HAM

```
In [ ]: HTML('<iframe width="560" height="315" src="https://www.youtube.com/embed/R13BD8qKeTg?
```