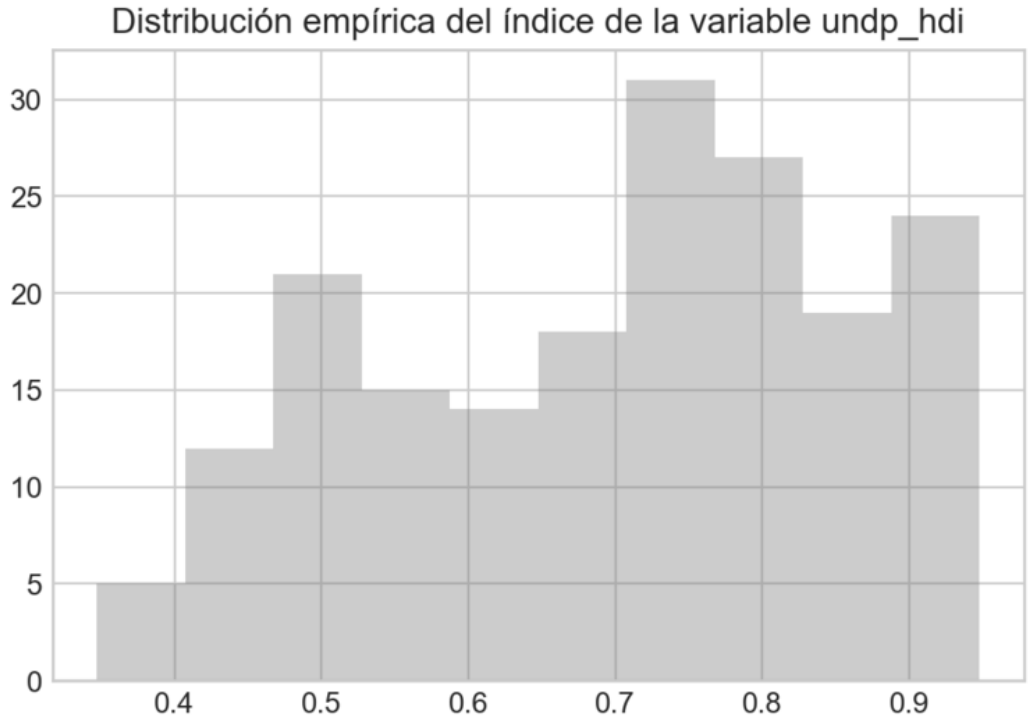
{desafío}
latam_

Variables Aleatorias y Gráficos _

Histogramas

Finalidad del histograma

- El histograma nos permite generar una **representación empírica** de una variable contínua.
- Se genera a partir de una serie de **casillas** (generalmente definidas por el programa) que cuenta la cantidad de ocurrencias entre cierto rango.
- El eje X representa el rango empírico de valores, el eje Y representa la frecuencia.



Matplotlib

- La librería de facto para implementar gráficos con Python se conoce como matplotlib, la cual está enfocada a reproducir gráficos estáticos en 2D. También se pueden realizar gráficos en 3D, aún cuando su soporte sea algo limitado.
- Podemos importar matplotlib de la siguiente manera:

```
import matplotlib.pyplot as plt
```

- Al importar este módulo, tendremos acceso a los componentes básicos, así como especificaciones de los elementos visuales.

Implementando el Histograma

- La implementación base del histograma se puede realizar con:

```
plt.hist(df['undp_hdi'])
```

- El método inferirá de forma automática cuál es la mejor representación de los bins (casillas) para esa variable.
- Dado que el eje X detalla el rango de valores que consideró, el eje Y nos informa sobre la cantidad de casos ocurrentes en el rango de valores.

Relevancia de los datos perdidos

- Una buena práctica es asegurarse que las variables no contengan datos perdidos.
- Para inspeccionar los datos perdidos de una variable podemos implementar la siguiente línea de código:

```
df['undp_hdi'].isnull()
```

- Dependiendo de la cantidad de datos perdidos, podemos ignorarlos o pensar en alguna estrategia de imputación.

Visualizando medias

- Si bien el histograma informa sobre la cantidad de ocurrencias, podemos hacerlo aún más informativo al incluir las medias.

```
hdi_drop = df[‘undp_hdi’].dropna()

plt.hist(hdi_drop)
plt.axvline(hdi_drop.mean(), color=’tomato’)
```

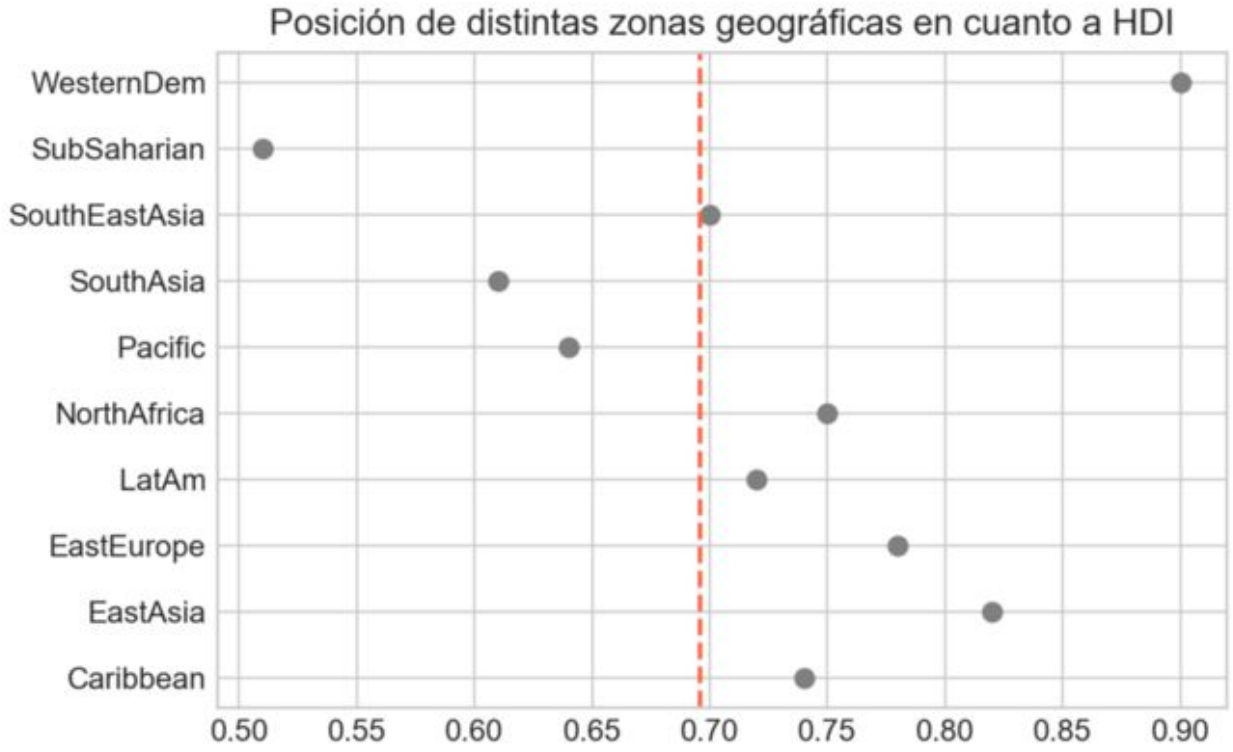
- La inclusión de medias en el gráfico nos permite evaluar cuántas observaciones se sitúan bajo o sobre la media global

Visualizando medias

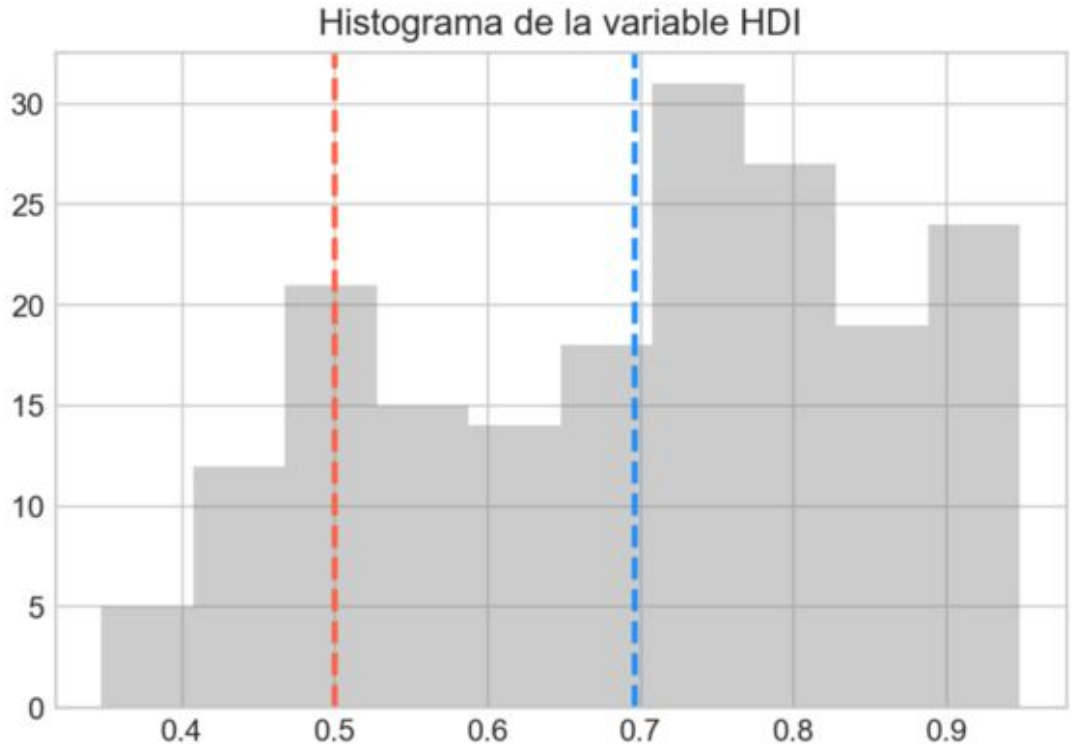
- Otra variante son los **dotplots**, que permiten desagregar el comportamiento de una variable, condicional a la pertenencia de las observaciones. Esto lo podemos lograr mediante

```
# con groupby logramos generar representaciones internas del dataframe
group_hdi_mean =
hdi_group.groupby('region_recod')['undp_hdi'].dropna().mean()
# dado que el retorno es una serie, podemos acceder a sus valores e
índices
plt.plot(group_hdi_mean.values, group_hdi_mean.index)
plt.axvline(hdi_dropna.mean())
```

Visualizando medias



Visualizando medias



{desafío} latam_

Variables Aleatorias

¿Qué es una variable aleatoria?

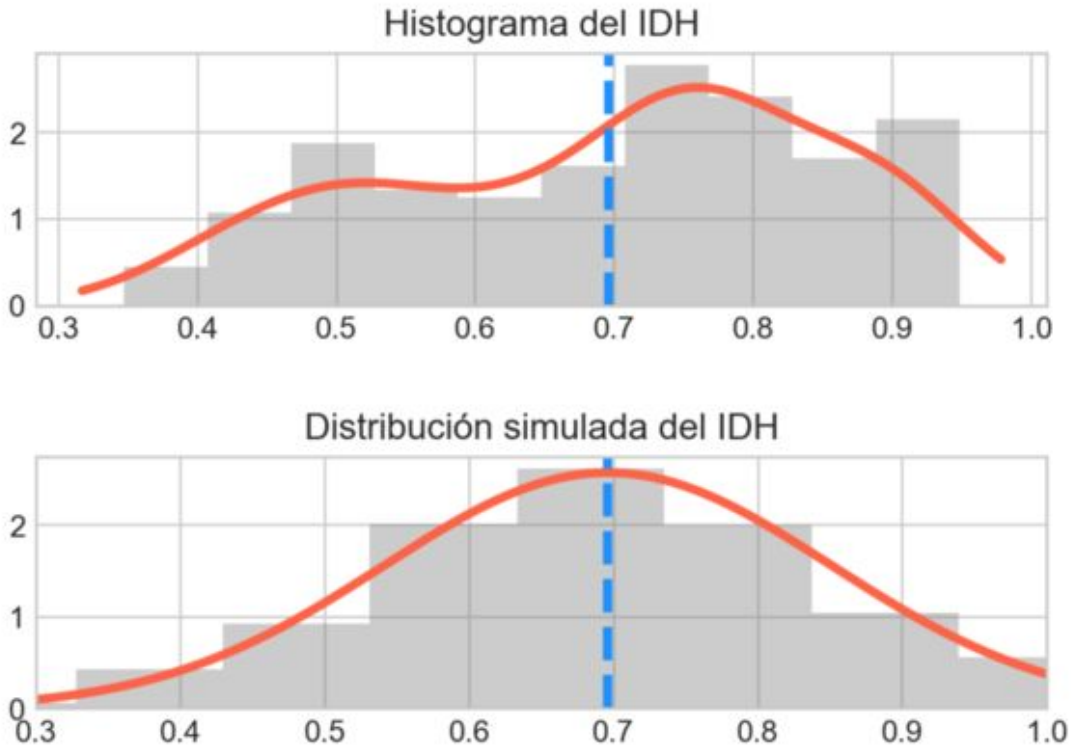
- Variable que toma un valor numérico único en un espacio muestral finito.
- Proveen una descripción sobre cómo se comporta un proceso de generación de datos.
- Hablamos de aleatoriedad cuando **no tenemos certeza sobre el comportamiento de cada evento específico**.
- Existen dos grandes familias de variables aleatorias:
  - Variables aleatorias continuas.
  - Variables aleatorias discretas.

Variables Aleatorias Continuas

Distribución Normal

- Permite aproximar una serie de fenómenos tales como altura, peso, coeficiente intelectual.
- Es probablemente una de las distribuciones más utilizadas.
- También presenta una serie de características deseables que facilitan el posterior trabajo de inferencia estadística
- Depende de dos parámetros:
  - Media
  - Varianza

Distribución Normal en el IDH

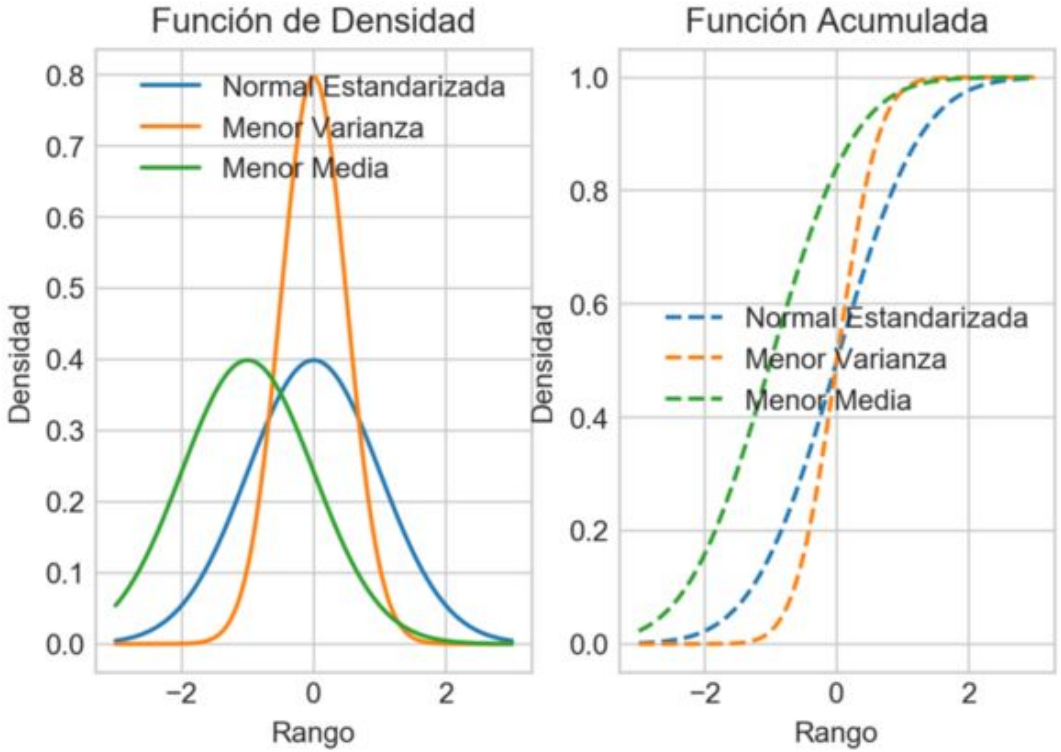


Componentes de la Distribución Normal

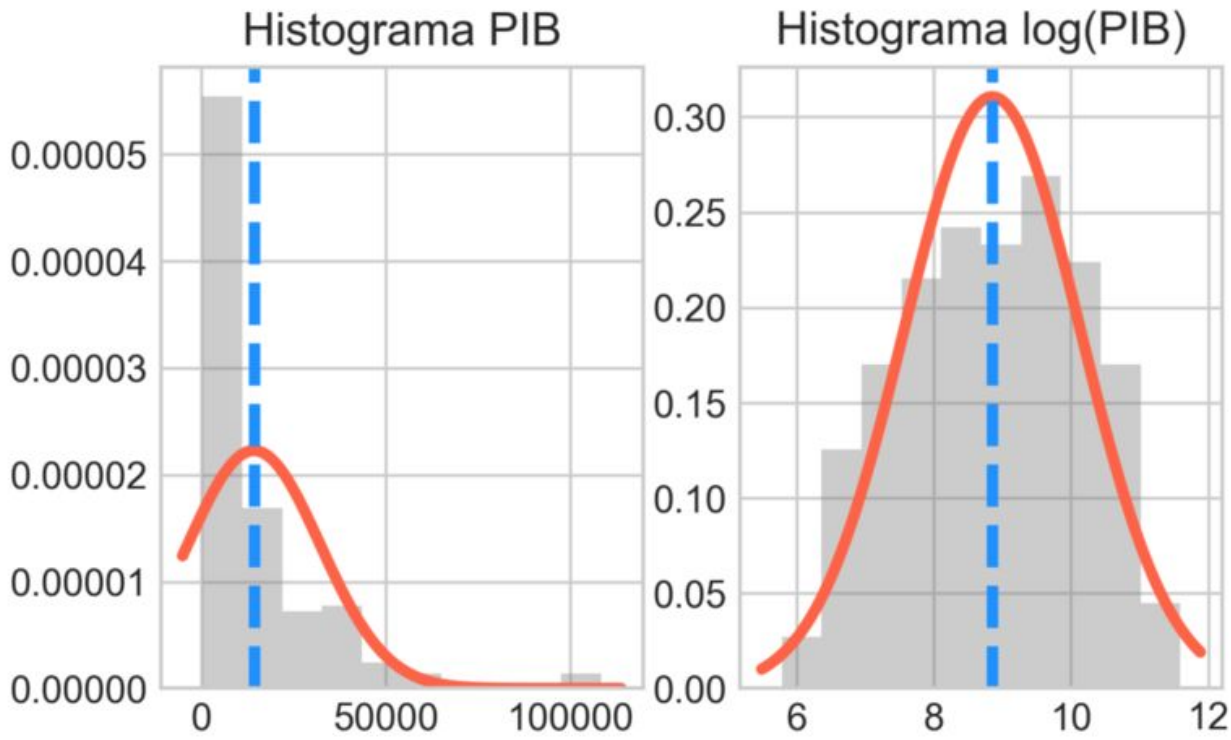
- Resulta que ya tenemos conocimiento sobre los primeros dos momentos de una variable.
- Con estos ya estamos en capacidad de aproximarnos al comportamiento de la variable.
- Son dos los componentes a tomar en cuenta:
  - Media
  - Varianza

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

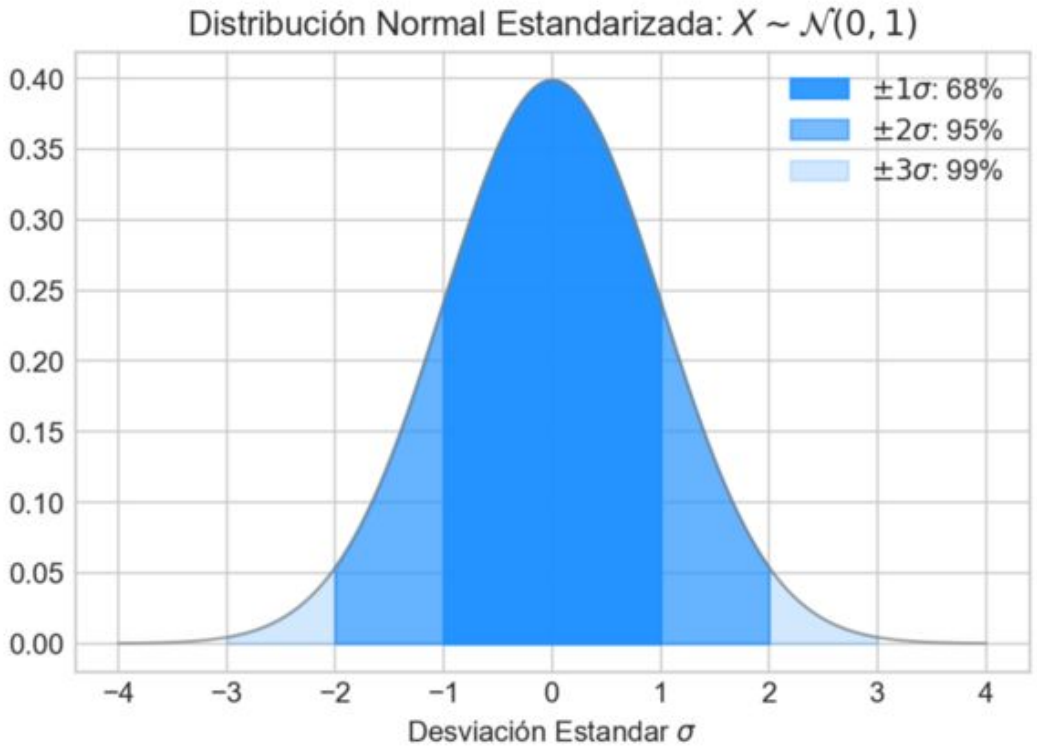
Comportamiento de la Distribución Normal



Posibles aproximaciones a la Normal



Distribución Normal Estandarizada



{desafío} latam_

Puntajes Z

- Permite analizar el comportamiento específico de una observación respecto a la media.

$$\text{Puntaje Z} = \frac{x_i - \bar{x}}{\sigma}$$

- Nos permite regularizar la comparación entre distintas variables.

Variables Aleatorias Discretas

Variables Aleatorias Discretas

- Nos permite resumir el comportamiento de un fenómeno discreto mediante el conteo.
- La distribución de una variable aleatoria X generalmente se especifica mediante el mapeo de todos los posibles valores y una probabilidad de masa:

$$p(x) = p_X(x) = \Pr[X = x]$$

Ensayo de Bernoulli

- Sabiendo que existen dos posibles eventos en un espacio muestral finito, **el ensayo de Bernoulli es la representación de una iteración del experimento en sí.**
- La distribución depende de un parámetro continuo $\mu \in [0, 1]$ que representa la probabilidad de ocurrencia.
- Podemos utilizar el módulo para obtener el comportamiento de un ensayo específico.

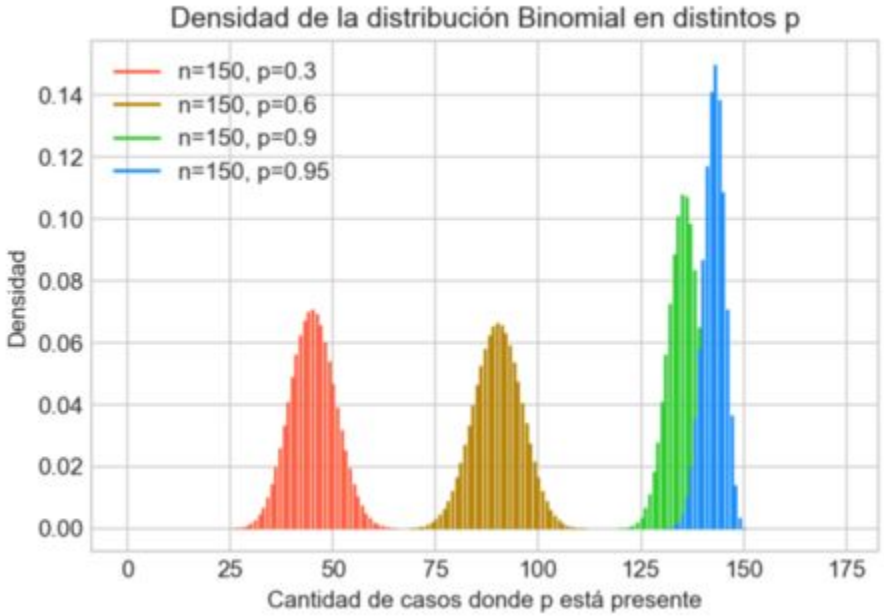
```
from scipy.stats import bernoulli
mu, sigma = bernoulli.stats(.65)
```

Distribución Binomial

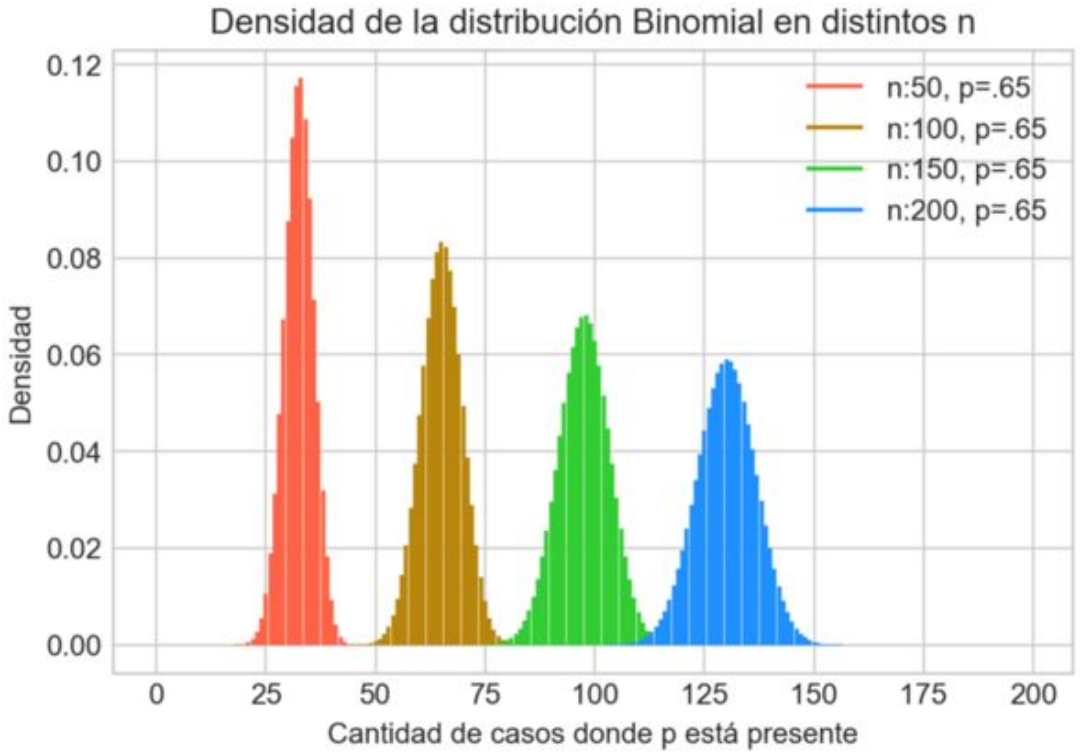
- Limitante del ensayo de Bernoulli: **ejemplifica el comportamiento de un caso cualquiera en una variable aleatoria.**
- Generalmente estamos interesados sobre la distribución de la tasa de éxito o fracaso de un fenómeno discreto.

```
from scipy.stats import binom
prob_democratic = binom(len(df), .654)
mu, sigma = prob_democratic.stats()
```

Comportamiento de la Distribución Binomial



Comportamiento de la Distribución Binomial



{desafío} latam_

Aspectos Asintóticos

¿Qué significa Asintótico?

- Resulta que el comportamiento de las variables aleatorias lo podemos generalizar al asumir un comportamiento asintótico.
- Para efectos prácticos del curso, cuando hablemos de comportamiento asintótico es **asumir que el tamaño de la muestra tiende al infinito.**

Ley de los Grandes Números

- En una sucesión infinitas de variables aleatorias i.i.d, el promedio de la sucesión será:

$$\bar{X} = (x_1 + x_2 + \cdots + X_n)/n$$

- La ley afecta la convergencia de la muestra respecto a la media poblacional

Teorema del Límite Central

- Si tenemos una secuencia de variables aleatorias i.i.d:

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{N}}} \xrightarrow{d} \mathcal{N}(0,1)$$

- Es una piedra angular la inferencia estadística dado que relaja los supuestos sobre la distribución de las variables.

{desafío} latam_

Academia de talentos digitales

www.desafiolatam.com