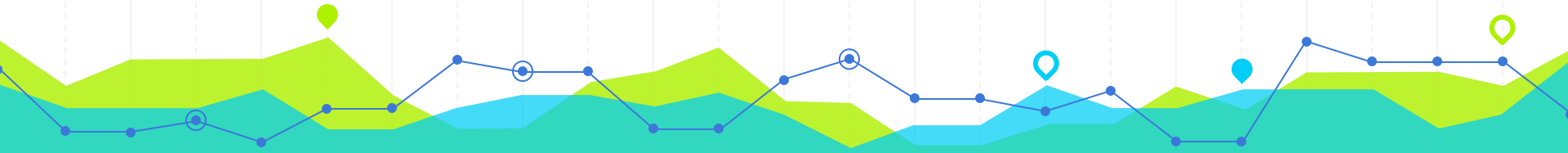
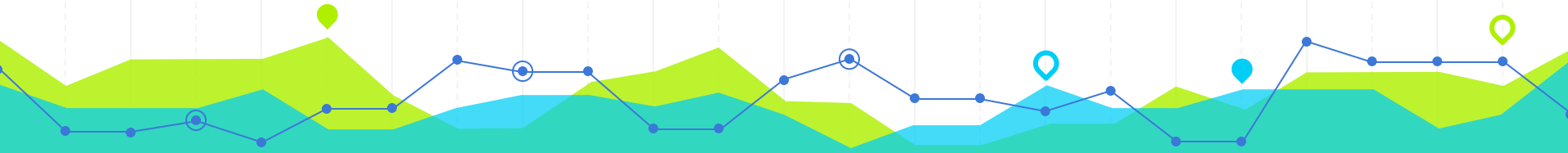
Universidad de Guanajuato
División de Ciencias Naturales y Exactas

BIOESTADÍSTICA

Luis Javier Torres Tetuan
Jesús Hernández González
Rodolfo Ferro Pérez



ANÁLISIS DE DATOS EN TWITTER



1 Introducción

¿Por qué hacer análisis de datos en redes sociales?

2. PLANTEAMIENTO DEL PROBLEMA

Pretendemos **analizar** qué tan necesario es para 2 usuarios elegidos al azar, usar el total de caracteres permitidos en la red social (**Twitter**) para causar un mayor **impacto** en lo que desean postear.

Los individuos en cuestión son un usuario especializado en **biología** y uno especializado en **política**.

Diferencias

entre un **tweet** de un **usuario** especializado en temas de **biología** y uno de **política**.

3. OBJETIVOS

- ***Análisis estadístico*** *de la información de un usuario con contenido específico* *(Biología)*

- ***Comparación estadística*** *con otro tipo de usuarios* *(Política)*

4. HIPÓTESIS

Las medias entre tweets de un usuario especializado en distintas áreas **son iguales**, es decir:

$$H_0: \mu_E = \mu_P = \mu$$

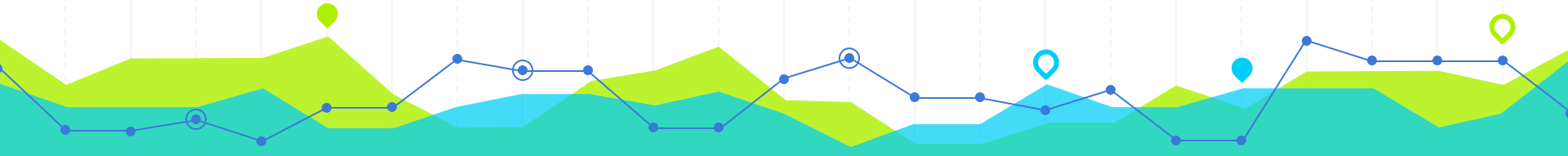
$$H_A: \mu_i \neq \mu_j, \quad \text{para } i \neq j, \quad i, j = \{E, P\}$$

Donde:

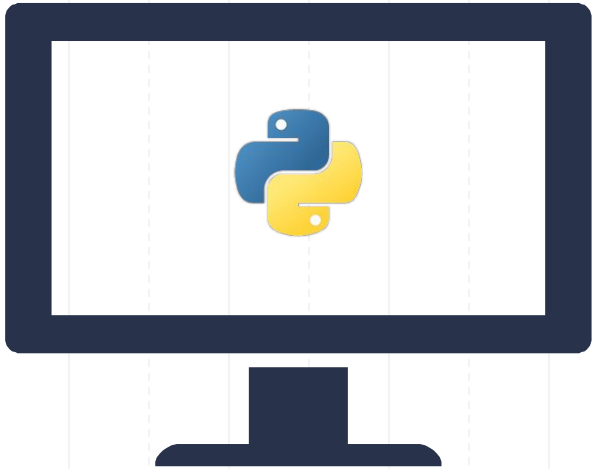
μ es la media global

μ_E es la media de usuario especializado

μ_P es la media de usuario promedio



5. Materiales:



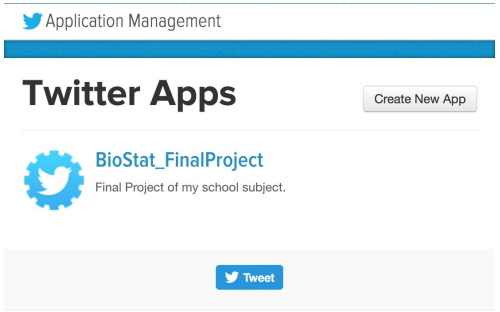
- Computadora
- Lenguaje de Programación

6. PROCEDIMIENTO

1. Creación de una aplicación en Twitter

2. Extracción de la información

3. Análisis de los datos




```
In [4]: for tweet in tweets:
            print(tweet.text)
            print("Fecha de creación:", tweet.created_at)
            print("Geolocalización:", tweet.geo)
            print("Longitud del tweet:", len(tweet.text), '\n')
```

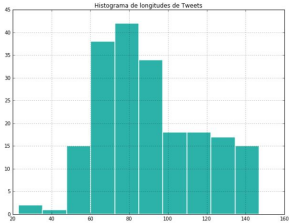
```
A cis cold memory element &amp;amp; a trans epigenome reader med
iate Polycomb silencing of FLC by vernalisation https://t.c
o/83utueE88G
Fecha de creación: 2016-11-29 17:26:04
Geolocalización: None
Longitud del tweet: 131

Cytokinin response factors integrate auxin and cytokinin pat
hways for female reproductive organ development https://t.c
o/JBN8QzZuIt
Fecha de creación: 2016-11-29 14:25:28
Geolocalización: None
Longitud del tweet: 131

Ovary-derived precursor gibberellin A9 is essential for fema
le flower development in cucumber https://t.co/VvCeCghuD8
Fecha de creación: 2016-11-29 14:25:10
Geolocalización: None
Longitud del tweet: 117
```

Histograma:

```
In [13]:
fig, ax = plt.subplots(figsize=(10,7.5))
plt.hist(data['Longitud'], normed=False,
color='lightseagreen', linewidth=1.5,
edgecolor='white')
plt.title("Histograma de longitudes de Tweets")
plt.grid(True)
```



7. ANÁLISIS DE DATOS

- El análisis lo hicimos con base en la longitud de texto escrito en cada tweet
- Los tweets usados fueron los 200 más recientes de cada usuario
- Los usuarios analizados fueron 2:
  @plantbiology
  @EPN

Extrajimos 200 tweets para cada usuario y los ordenamos en una tabla de este tipo.

	Tweet	Longitud
0	A cis cold memory element & a trans epigen...	131
1	Cytokinin response factors integrate auxin and...	131
2	Ovary-derived precursor gibberellin A9 is esse...	117
3	Demethylation of ERECTA receptor genes by IBM1...	119
4	5000-year-old cobs reveal corn domestication i...	79

7.1. RESULTADOS MATEMÁTICOS

MEDIDAS DE TENDENCIA CENTRAL:

Media aritmética:

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

In [8]:

```
np.mean(data['Longitud'])
```

Out[8]:

90.6

Mediana:

$$M_{pos} = \frac{n+1}{2}$$

In [10]:

```
np.median(data['Longitud'])
```

Out[10]:

85.5

Moda

In [9]:

```
np.argmax(np.bincount(data['Longitud'])
```

Out[9]:

83

7.1. RESULTADOS MATEMÁTICOS

MEDIDAS DE TENDENCIA CENTRAL:

Media armónica:

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

In [12]:

```
len(data['Longitud'])/np.sum(1./data['Longitud'])
```

Out[12]:

```
82.61347934524194
```

Media geométrica:

$$G = \sqrt[n]{x_1 \cdots x_n}$$

In [11]:

```
np.prod(np.power(data['Longitud'], 1./len(data['Longitud'])))
```

Out[11]:

```
86.765131610856642
```

7.1. RESULTADOS MATEMÁTICOS

MEDICIÓN DE LA VARIABILIDAD:

Desviación estándar:

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

In [17]:

```
np.std(data['Longitud'])
```

Out[17]:

```
25.976912826585078
```

Varianza:

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

In [18]:

```
np.var(data['Longitud'])
```

Out[18]:

```
674.8000000000003
```

7.1. RESULTADOS MATEMÁTICOS

MEDICIÓN DE LA VARIABILIDAD:

Grados de libertad:

$$\gamma = N - 1$$

In [19]:

```
len(data['Longitud']) - 1
```

Out[19]:

199

Coeficiente de variación:

$$CV = \frac{\sigma}{\mu} \times 100$$

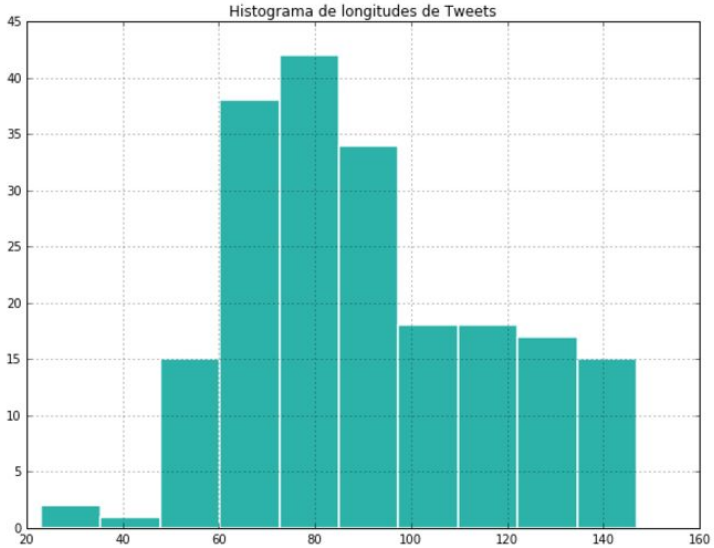
In [20]:

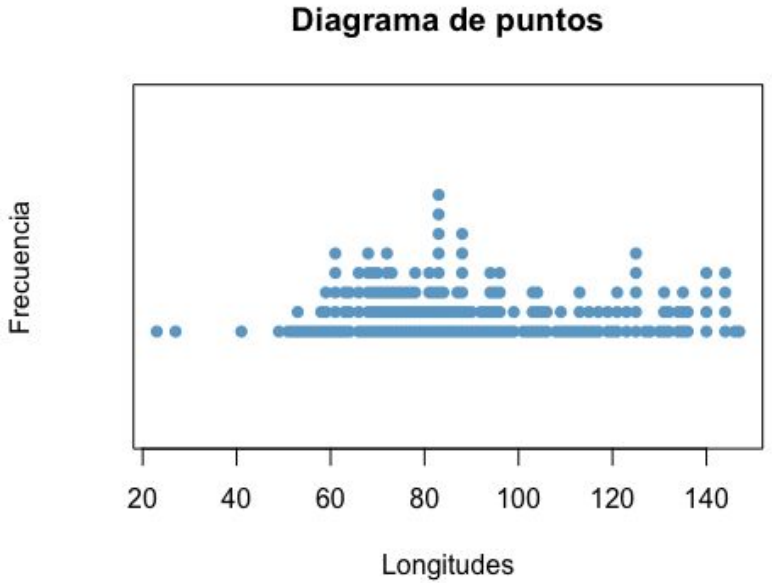
```
np.std(data['Longitud'])/np.mean(data['
gitud'])*100
```

Out[20]:

28.672089212566316

7.2 RESULTADOS GRÁFICOS



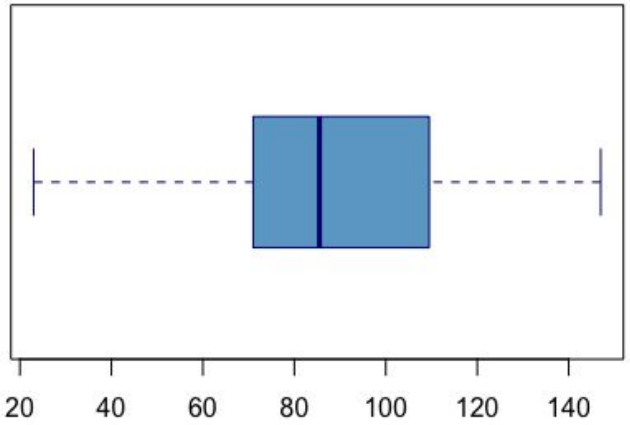


7.2 RESULTADOS GRÁFICOS

Diagrama de tallo y hoja

```
 2 | 3 7
 4 | 1 9
 5 | 1 2 3 3 4 5 6 7 8 8 9 9 9
 6 | 0 1 1 1 1 1 2 3 3 3 4 4 4 6 6 6 6 7 8 8 8 8 8 9 9 9 9
 7 | 0 0 0 0 1 1 1 2 2 2 2 2 3 3 3 3 4 4 4 5 5 5 6 6 6 7 7 7 8 8 8 8 9 9
 8 | 0 0 1 1 1 1 2 2 2 3 3 3 3 3 3 3 4 4 4 5 5 6 6 7 7 7 8 8 8 8 8 8 9 9
 9 | 0 0 1 2 2 3 3 4 4 4 4 5 5 5 6 6 6 6 7 8 9 9
10 | 1 2 3 3 3 4 4 4 5 5 6 6 8 9 9
11 | 0 1 2 3 3 3 4 5 5 6 7 7 9 9
12 | 0 1 1 1 3 3 5 5 5 5 5 7 8
13 | 0 1 1 1 2 2 4 4 5 5 5 6 6
14 | 0 0 0 0 4 4 4 4 6 7
```

Diagrama de caja y bigotes




```
> quantile(x)
    0%    25%    50%    75%   100%
23.00  71.00  85.50 109.25 147.00
```

7.3 COMPARACIÓN DE MEDIAS

Comparación de medias (desv. est.):

Hipótesis:

$$H_0 : \bar{x_1} = \bar{x_2}$$
$$H_A : \bar{x_1} \neq \bar{x_2}$$

$$F = \frac{s_1}{s_2}, F \geq 1$$

In [28]:
```
F = np.std(data['Longitud']) / np.std(p_data['Longitud'])
print("F = {} > 1".format(F))
```

```
F = 1.708643605642507 > 1
```

Entonces,

$$F_0 = 1.7086,$$
$$GL_1 = GL_2 = 199.$$

F_0	**Comp.**	$F_{0.05,199,199}$	H_0
1.7086	>	1.26	X

Se rechaza la hipótesis de tener la misma desv. estándar.

7.3 COMPARACIÓN DE MEDIAS

Ahora se calculan los estadísticos:

$$t = \frac{(\bar{x_1} - \bar{x_2})}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$GL = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}\right)}$$

In [37]:
```
t = (np.mean(data['Longitud']) - np.mean(p_data['Longitud'])) / (np.sqrt((n
print("t = {}".format(t))
print("|t| = {}".format(np.abs(t)))
```

```
t = -16.88665910892693
|t| = 16.88665910892693
```

In [42]:
```
GL = (((np.std(data['Longitud'])**2/200) + (np.std(p_data['Longitud'])**2/2
print("GL = {}".format(GL))
```

```
GL = 321.011345864429
```

7.3 COMPARACIÓN DE MEDIAS

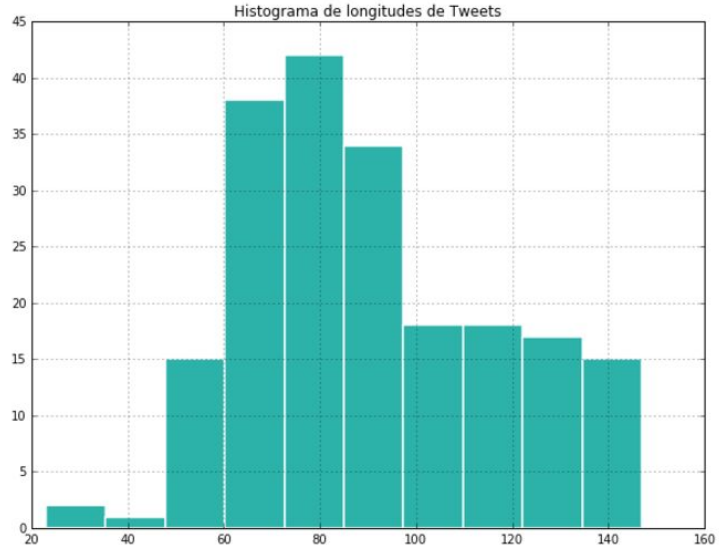
Se hace la comparación:

t_0	**Comp.**	$t_{0.05,321}$	H_0
16.8866	>	1.96	X

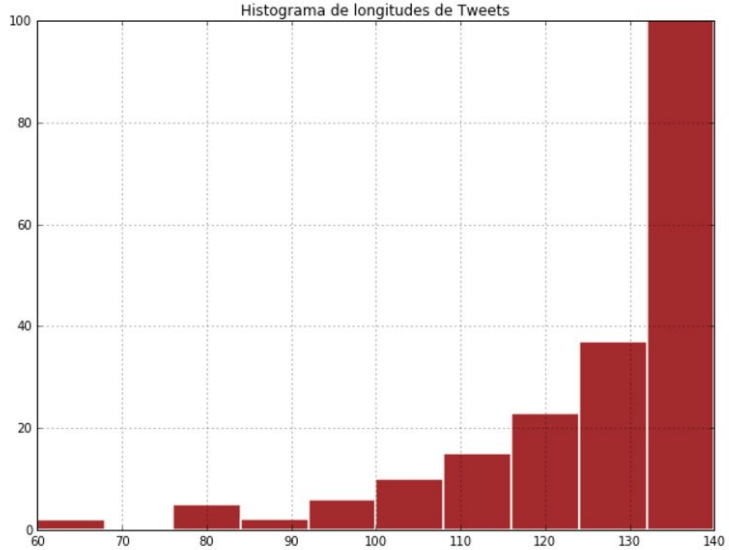
Y de aquí se concluye que SÍ hay diferencia estadística entre las medias.

7.3 COMPARACIÓN DE MEDIAS

@plantbiology



@EPN



7.3 COMPARACIÓN DE MEDIAS

Comparación de medias (*t* pareada):

Hipótesis:

$$H_0 : \mu_d = 0$$

Calculamos el estadístico:

$$t = \frac{\bar{d}\sqrt{n}}{s_d} = 18$$

$$GL = 200 - 1 = 199$$

Hacemos la comparación:

t_0	**Comp.**	$t_{0.05,199}$	H_0
18	>	1.96	X

De aquí que hay consistencia en la diferencia estadística de las medias.

8. DISCUSIÓN

Plant biology

Con una media de 90.6 caracteres por tweet, de una muestra de 200 tweets.

EnriquePeñaNieto

Con una media de 125.6 caracteres por tweet, en una muestra de 200 tweets.

140

Es el número máximo de caracteres que están disponibles para cada Tweet escrito por usuario.

8. DISCUSIÓN

Después de analizar los doscientos tweets de cada usuario, se puede estimar la **media** en su uso de caracteres, el límite se conoce en 140 y se nota que el usuario EPN utiliza en la mayoría de sus tweets el **máximo** posible, ya que trata de generar empatía y sentido de profesionalismo.

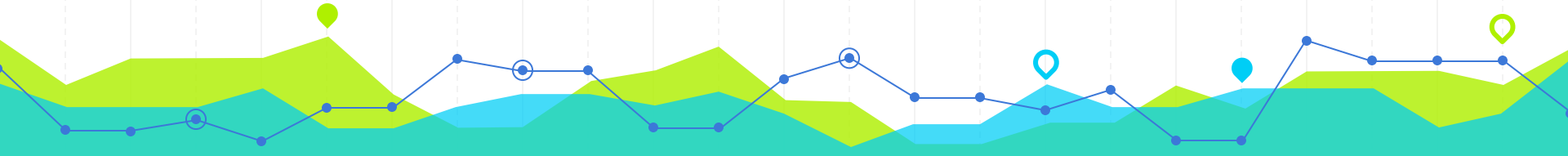
8. DISCUSIÓN

En cambio en el usuario especializado en biología Plant Biology busca ser **conciso, claro y directo** a la hora de postear para atraer la atención y expresar en **pocas palabras** la idea principal de un tema específico que usualmente profundiza dejando un link de acceso.

9. CONCLUSIONES

Los valores estadísticos para el uso de caracteres por parte del usuario especializado en **biología son menores** comparación a un usuario de índole **política**, tal vez se debe al ámbito en el que debe expresar cada usuario.

El político tiene a adornar sus tweets para aumentar sus seguidores, en cambio un usuario especializado sabe que debe **ser claro** y expresar en el **menor número de palabras** una idea principal para atraer atención del público en general



10. BIBLIOGRAFÍA

- Miller, *Estadística y Quimiometría para Química Analítica*, 2005.
- Jupyter Notebook Documentation: https://jupyter.readthedocs.io/en/latest/
- Tweepy Documentation: http://docs.tweepy.org/en/v3.5.0/

¿Alguna pregunta?

CRÉDITOS:

- Rodolfo Ferro Pérez
- Luis Xavier Torres Tetuan
- Jesús Hernández González

https://rodolfoferro.github.io/biostat_finalproj/