

Departamento de Computación, FCEyN, UBA

Procesamiento del Habla

Agustín Gravano

1er Cuatrimestre 2017

Estadística Aplicada y Lingüística Empírica



Antes de empezar...

- Ir a <http://tinyurl.com/experimento-ITH>
- Ingresar tu altura en el casillero correspondiente.
- Ingresar la altura de algún amigo/a del sexo opuesto y mayor de 18 años, que no esté hoy acá.
- En **centímetros** (ej: 1.80m = 180cm)

Ciencias Empíricas

- Ciencias Naturales
 - Biología, Química, Geología, Astronomía, ...
- Ciencias Sociales
 - Sociología, Economía, Lingüística, ...
- Método Empírico
 - Observaciones del mundo + (a veces) experimentación.
 - Inferencias, generalizaciones.

Hipótesis

“En promedio, el hombre es más alto que la mujer.”

¿Cómo hacemos para evaluar esta hipótesis?

- Para estar seguros: medir la altura de todos los seres humanos del planeta, promediar y comparar.
- Alternativa: obtener una muestra representativa, medir, promediar y comparar.

Población y Muestra

- **Población**
 - Colección de *todos* los resultados, respuestas, medidas o cantidades de interés.
- **Muestra**
 - Subconjunto de la población.
- **Ejemplo (datos ficticios):**
 - *En una encuesta, se preguntó a 250 alumnos de la UBA si tienen cuenta en redes sociales. 195 respondieron que sí.*
 - **Población:** Respuestas de todos los alumnos de la UBA.
 - **Muestra:** Respuestas de los alumnos de la encuesta.

Nuestro ejemplo

- Hipótesis:
 - *“En promedio, el hombre es más alto que la mujer.”*
- Población:
 - Todos los seres humanos del planeta.
 - ¿realmente todos? ¿bebés y niños? ¿amputados?
- Muestra:
 - Para ser representativa, debería tener individuos de todas las regiones, etnias, etc.

Nuestro ejemplo

- Hipótesis: (menos ambiciosa)
 - *“En promedio, el hombre es más alto que la mujer en el área metropolitana de Buenos Aires, teniendo en cuenta individuos 'sanos' mayores a 18 años.”*

Hombre - Altura en cm	Mujer - Altura en cm
172	160
175	160
183	165
179	165
182	183
192	165
191	172
183	170
172	165
168	165
175	163
178	153
172	170
190	169
182	160
166	156
180	165
168	164
170	165
183	173
181	160
180	185
180	165
188	173
188	165

mean(hombre) = 176.8 cm
mean(mujer) = 162.25 cm

(Datos de una cursada anterior.)

Gráfico de barras (bar plot).
Muy poco informativo.

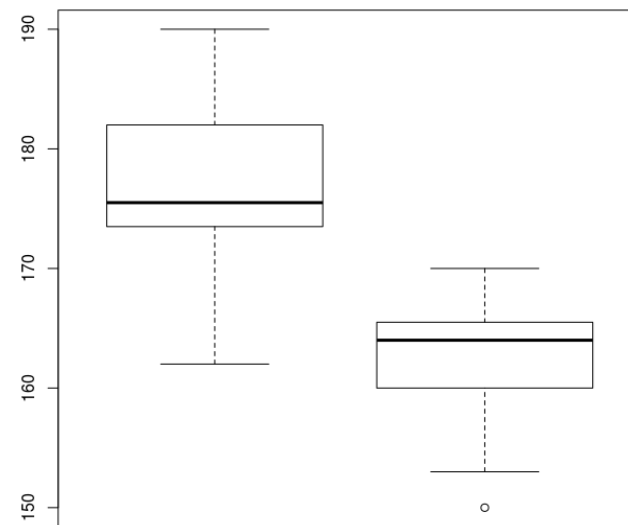
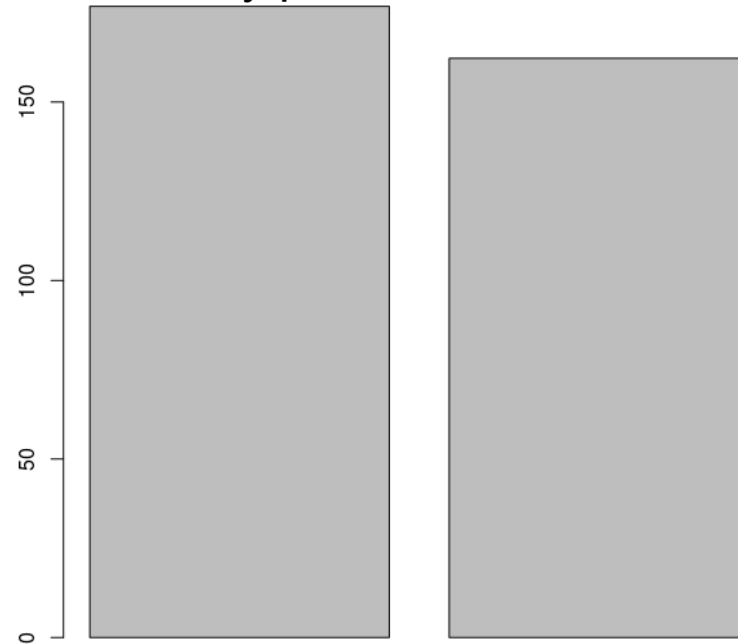


Gráfico de cajas (box plot).
Bastante mejor.

Analizando los resultados...

- ¿Cuán confiables son los resultados?
- ¿Podemos generalizarlos a la población?
- ¿Qué chances hay de que los resultados sean consecuencia del azar?
- ¿Cuán replicable es el experimento?
- ¿Los resultados son **significativos**?
- Respuesta: tests estadísticos.

Hipótesis nula vs. alternativa

- Hipótesis nula (H_0): $\mu_H = \mu_M$
 - No existen diferencias. (Por defecto suponemos esto.)
- Hipótesis alternativa (H_1): $\mu_H > \mu_M$
 - Sí hay diferencias. (Pero necesitamos evidencia para sostenerlo.)
- Test estadístico:
 - ¿Tenemos suficiente evidencia para **rechazar** H_0 ?
 - Student's **t-test**: compara la media de dos muestras.

t-test de Student

- Input: dos grupos de muestras.
 - Muestras independientes.
 - Poblaciones con distribución Normal y misma varianza.
- Output: p -valor $[0,1]$ (entre otras cosas)
 - Es la probabilidad de observar un resultado como el obtenido (o más extremo), suponiendo que H_0 es cierta.
 - Si el p -valor es muy pequeño podemos *rechazar H_0*
 - Suele usarse $p < 0.05$ o $p < 0.01$ para determinar si un resultado es *estadísticamente significativo* (o sea, 5% o 1% chances de que sea erróneo concluir que existen diferencias).
 - **OJO:** “significativo” no es que la diferencia sea **grande**, sino que es **representativa** de lo que ocurre a nivel poblacional.

En Python...

```
Hombres = [175,173,187,165,180,180,176,168,183,175,170,184,174,  
175,175,181,185,162,190,178]
```

```
Mujeres = [162,154,155,167,165,167,166,165,170,164,170,165,163,  
160,153,160,160,164,165,150]
```

```
from scipy import stats
```

```
stats.ttest_ind(hombres, mujeres).pvalue / 2
```

Esta función calcula el t-test de dos colas.
Dividimos por 2 para pasarlo a una cola.

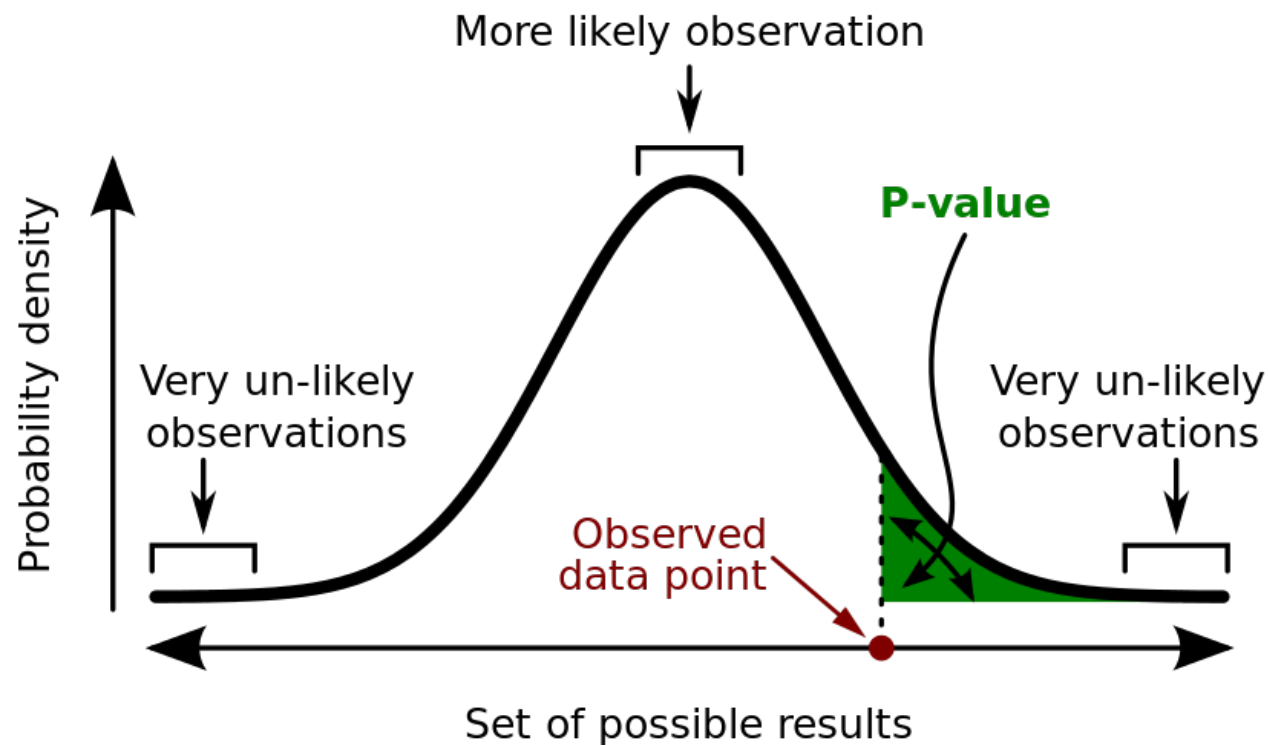
Resultado: 8.1214e-09

- ¿Cómo afecta la cantidad de datos a la significancia de los resultados?
- ¿Qué pasaba si hoy venían pocos alumnos?

(ver altura-por-genero.ipynb)

P-valor




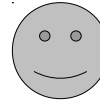




$P = \Pr(\text{observación} \mid \text{hipótesis}) \neq \Pr(\text{hipótesis} \mid \text{observación}) !!!$



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

t-test apareado

- Otra opción: que los datos vengan **de a pares**.
- Ejemplo:
 - *A un grupo de pacientes se les mide la temperatura antes de darles un medicamento, y una hora después. Pregunta: ¿les baja la temperatura?*

								
Antes:	x11	x21	x31	x41	x51	x61	x71	x81
Después:	x12	x22	x32	x42	x52	x62	x72	x82

- Paired t-test: Considera **pares** de muestras, en lugar de muestras totalmente independientes.
- En Python: `scipy.stats.ttest_rel`

Correlación: Ejemplo 1

- X = Consumo anual per cápita de cigarrillos (hombres, EEUU)

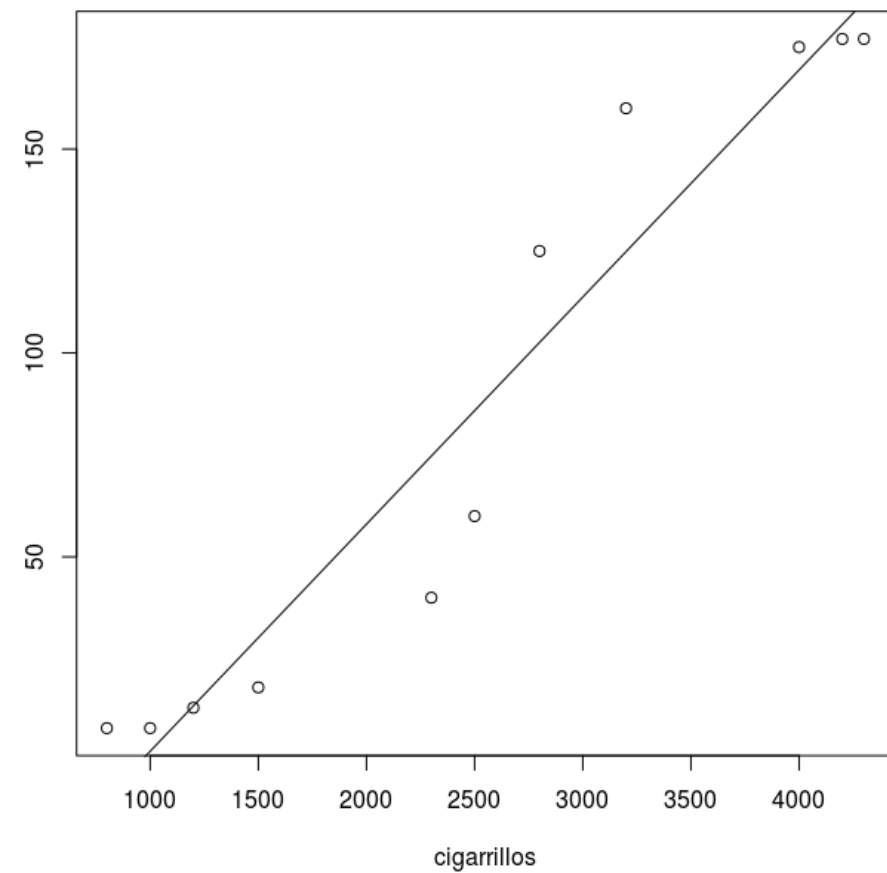
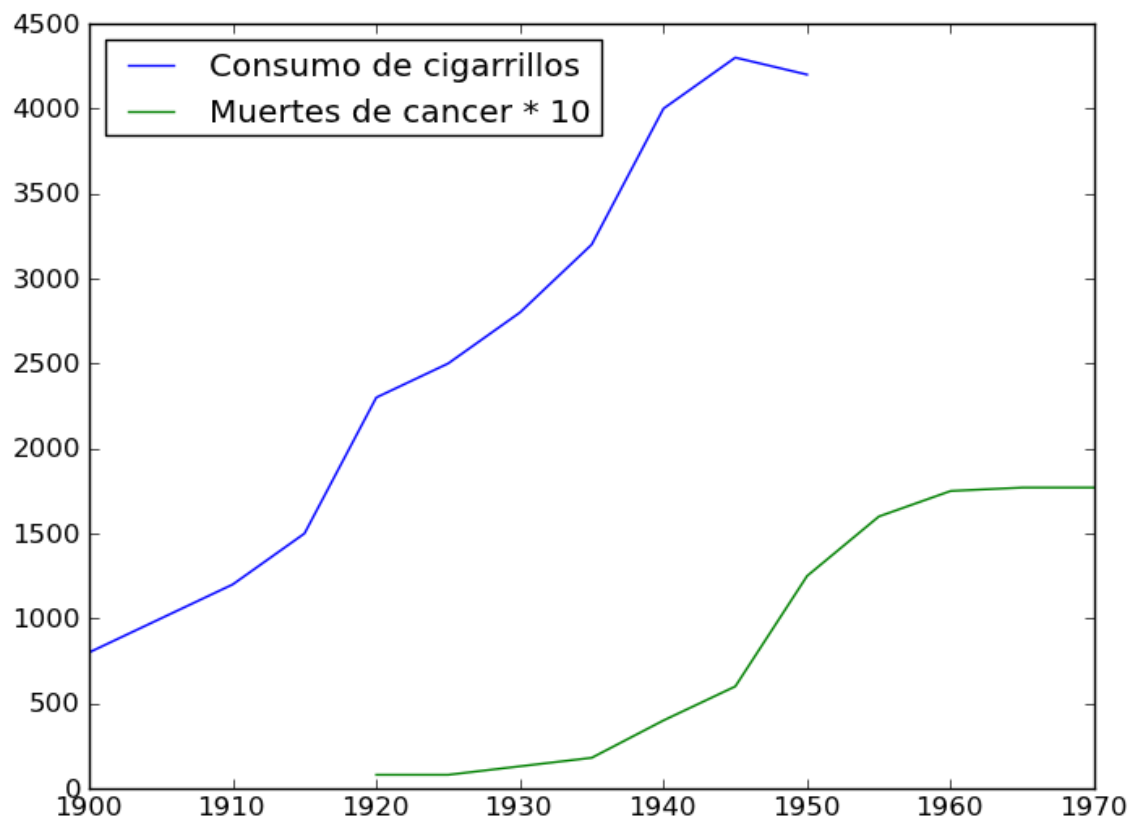
	1900	1905	1910	1915	1920	1925	1930	1935	1940	1945	1950
X	800	1000	1200	1500	2300	2500	2800	3200	4000	4300	4200

- Y = Muertes por cáncer de pulmón por cada 100.000 habitantes (hombres, EEUU)

	1920	1925	1930	1935	1940	1945	1950	1955	1960	1965	1970
Y	8	8	13	18	40	60	125	160	175	177	177

- ¿Cuán fuertemente correlacionadas están X e Y ?
- *cigarrillos-vs-cancer.ipynb*

(Basado en datos reales. Fuente: NIH)



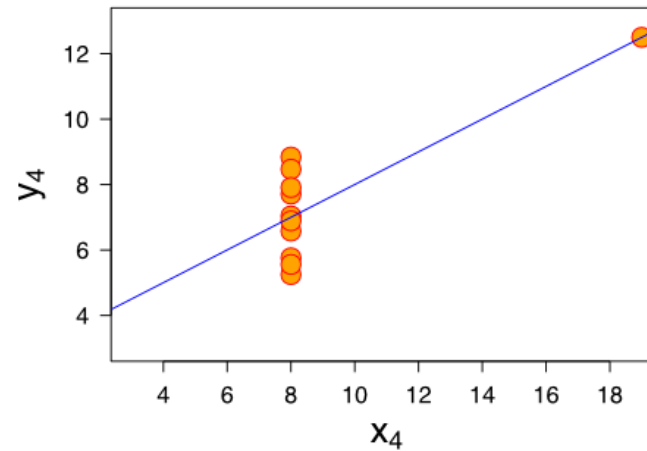
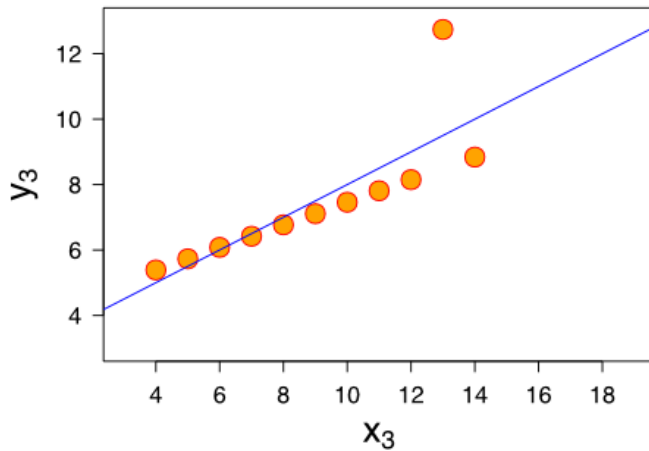
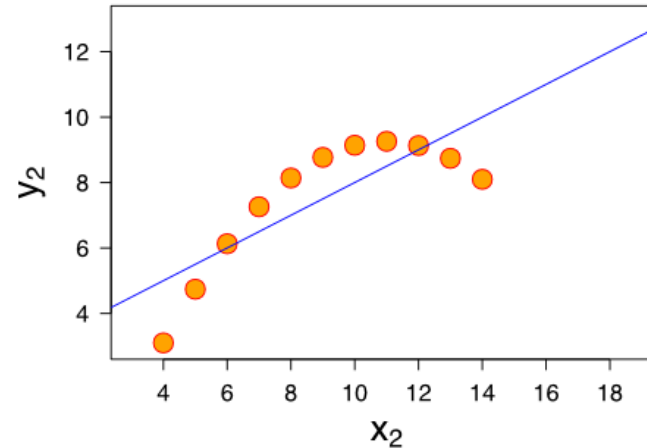
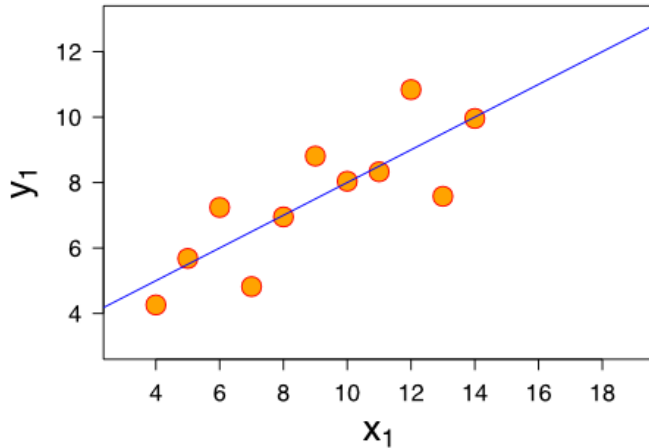
```
cigarrillos = [800,1000,1200,1500,2300,2500,2800,3200,4000,4300,4200]
cancer = [8,8,13,18,40,60,125,160,175,177,177]
```

```
stats.pearsonr(cigarrillos, cancer)[0]
```

Resultado: **0.961994**

```
plt.plot(cigarrillos, cancer, linestyle='', marker='.')
ajuste = np.poly1d(np.polyfit(cigarrillos, cancer, 1))
plt.plot(cigarrillos, ajuste(cigarrillos))
```

Correlación: Patologías



En todos los casos, $cor = 0.816$.

¡Siempre hay que visualizar!

Correlación: Ejemplo 2

- X = Consumo per cápita de helado en Bs.As. (en gramos).

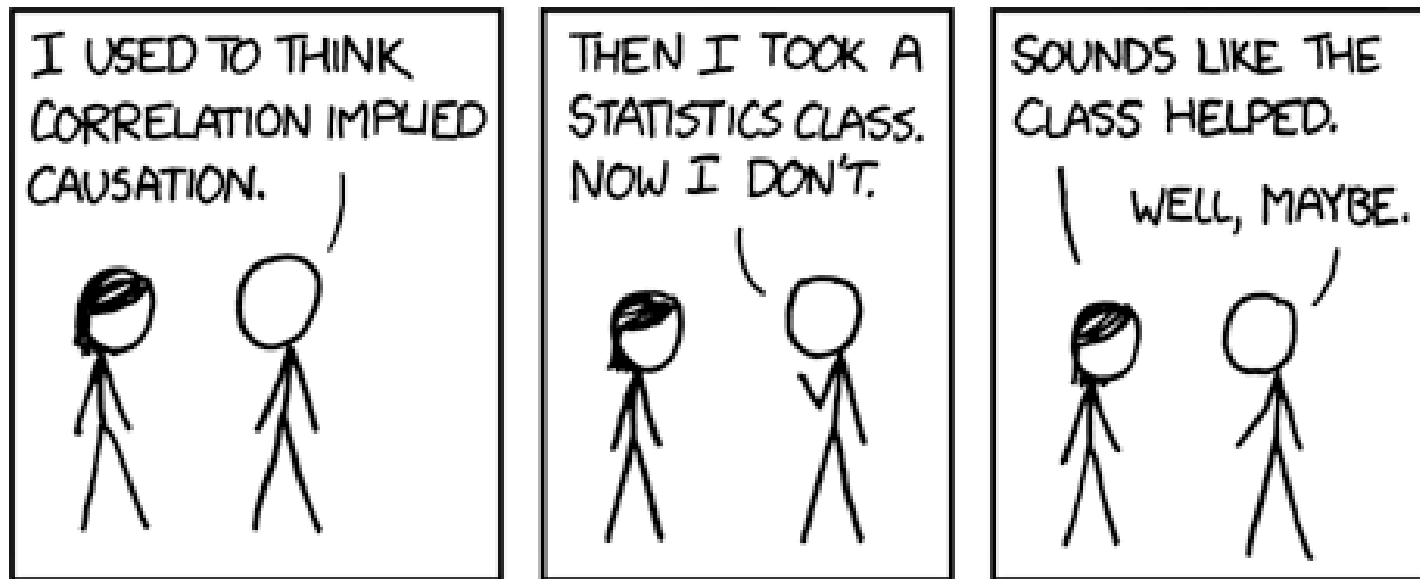
	Ene	Feb	Mar	Abr	May	Jun	Jul	Ago	Sep	Oct	Nov	Dec
X	1200	1150	900	540	300	280	540	270	430	650	730	1140

- Y = Ahogados en Bs.As. por cada 100.000 habitantes.

	Ene	Feb	Mar	Abr	May	Jun	Jul	Ago	Sep	Oct	Nov	Dec
Y	30	37	25	7	4.1	3.6	7.4	5.3	7.5	9	14	21

- ¿Cuán fuertemente correlacionadas están X e Y ?
- *helados-vs-ahogados.ipynb*

(Datos inventados.)



<http://xkcd.com/552/>

Spurious correlations:
<http://www.tylervigen.com>

¿Cómo podemos buscar
evidencia de causalidad?

Experimento controlado

- 1) Tomar N personas al azar, todos no fumadores.
 - Grupo F: N/2 personas → Que empiecen a fumar.
 - Grupo NF: N/2 personas → Que sigan sin fumar.
- 2) 20 años después, contar cuántos murieron de cáncer en cada grupo, y comparar (p.ej. usando el test de χ^2 : `chisq.test`).

	Cáncer	No Cáncer
Grupo F		
Grupo NF		

- **Ética!!!!**
- Estos estudios **sí** se hacen para testear hipótesis menos tremendas: remedios, marketing, biología, psicología, etc.

Lingüística Empírica

- **Hipótesis sobre el lenguaje.**
 - Sintaxis, semántica, fonología, prosodia, ...
- Estudios observacionales:
 - Ej: *Las frases inconclusas suelen terminar en H-L% o L-H%.*
 - Recolección y anotación de un **cuerpo de datos** (textos o audio).
 - Muestras = Cuerpos de datos. Población = ?
 - **Tests estadísticos** para estudiar las hipótesis.
- Estudios experimentales:
 - Ej: *Las frases que terminan en H-L% o L-H% suelen **percibirse** como inconclusas.*
 - Diseñar un experimento.
 - Factores (o tratamiento) vs. efectos (o variables de interés).

Lingüística Empírica: Ejemplo

- Hipótesis:
 - *“La voz del hombre es más grave que la de la mujer”*
- Procedimiento:
 - 1) Diseño del estudio
 - 2) Datos
 - 3) Observaciones/mediciones
 - 4) Análisis exploratorio de los datos (ej: visualizaciones)
 - 5) Análisis estadístico
 - 6) Conclusiones

Lingüística Empírica: Ejemplo

“La voz del hombre es más grave que la de la mujer”

1) Diseño del estudio

- Hipótesis nula (H_0) = El nivel tonal es **similar** en hombres y mujeres.
- Hipótesis alternativa (H_1) = El nivel tonal es **menor** en los hombres que en las mujeres.
- Vamos a realizar **observaciones**.
 - Si son **consistentes** con H_0 , decimos que “no encontramos evidencia contraria a H_0 ”.
 - Si son **inconsistentes** con H_0 , decimos que “tenemos evidencia para rechazar H_0 ”, y por ende para aceptar H_1 .

Lingüística Empírica: Ejemplo

“La voz del hombre es más grave que la de la mujer”



2) Datos

- Grabaciones de habla espontánea y leída realizadas por alumnos de la materia en 2011 para un TP.



- Espontánea: describir el tiempo.
- Leída: leer un texto del mismo tema.

El clima de la ciudad de Buenos Aires es templado pampeano húmedo. Considerando el período 1961-1990, normalmente empleado para designar los promedios climáticos, la temperatura media es de 17,6°C y la precipitación anual es de 1146 mm. A lo largo del siglo XX las temperaturas de la ciudad han aumentado considerablemente debido a la isla de calor (desarrollo urbano), siendo actualmente 2°C superior al de regiones cercanas mucho menos urbanizadas. Fundamentalmente las temperaturas nocturnas son las que han aumentado, lo que en verano suele dificultar el descanso nocturno de los porteños. Las precipitaciones también se han acrecentado desde 1973, como ya ocurrió en el anterior hemicycle húmedo: 1870 a 1920.

Ciudad de Buenos Aires

Lunes		Temperatura: 23° Humedad: 40% Vientos leves del Este
Martes		Temperatura: 20° Humedad: 60% Vientos moderados del Este

Río Gallegos

Lunes		Temperatura: 22° Humedad: 70% Vientos moderados a fuertes del Sudoeste
Martes		Temperatura: 16° Humedad: 90% Vientos fuertes del Sur

- 184 audios, grabados por 92 personas (46m, 46f; edad: 20-74) hablantes nativas de español, que vivían en Buenos Aires y alrededores.
- Wav mono, 16 kHz, 16 bits. Grabados con micrófonos no profesionales.
- /home/ph-30/clase-05/datos/NNNgMMh.{wav,ipu}
NNN=id hablante (000-092) g=f/m MM=edad h=r/s (habla leída o espontánea)
wav=audio ipu=transcripción de unidades sin pausas

Lingüística Empírica: Ejemplo

“La voz del hombre es más grave que la de la mujer”

3) Observaciones/mediciones

- Variable: media de la frecuencia fundamental (F0) en todo el archivo
 - Rango 75-300 Hz para hombres y 100-500 Hz para mujeres.
- Usamos el archivo [acoustics.praat](#) que ya vimos para extraer atributos acústicos.
- La función [run_praat](#) es un wrapper en Python.

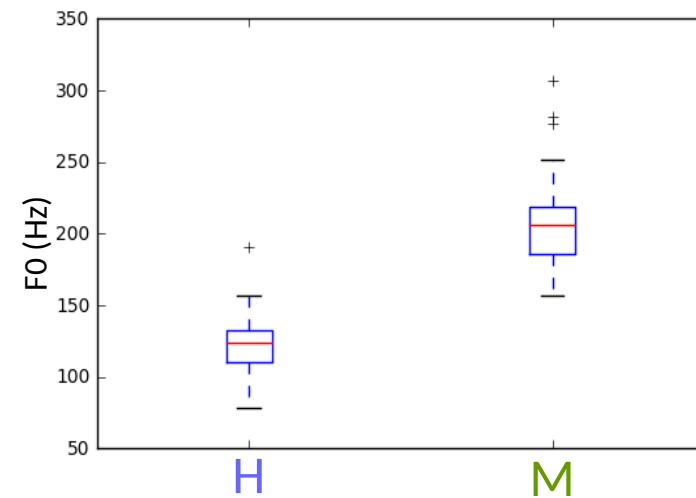
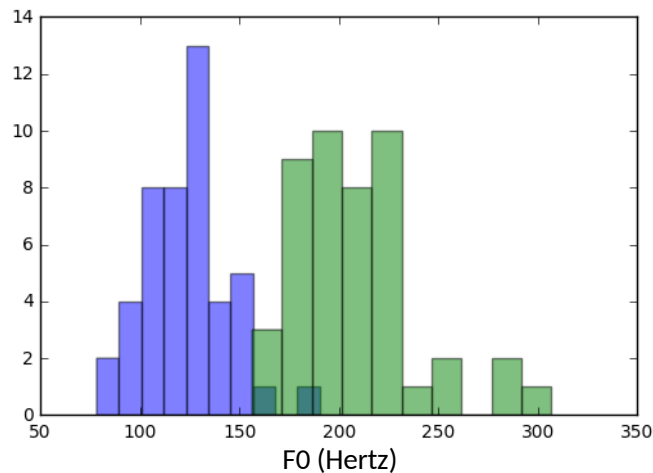
Lingüística Empírica: Ejemplo

“La voz del hombre es más grave que la de la mujer”

4) Análisis exploratorio de los datos

Hombres: 123.2 +/- 20.5 Hz

Mujeres: 207.7 +/- 30.3 Hz



niveltonal-por-genero.ipynb

Lingüística Empírica: Ejemplo

“La voz del hombre es más grave que la de la mujer”

5) Análisis estadístico

- One-tailed t -test de Student.
- Resultado: $p \approx 0$

6) Conclusiones

- El p -valor resultante es bajísimo, lo cual constituye sólida evidencia empírica de que deberíamos rechazar H_0 y quedarnos con H_1 : que los hombres en general tienen la voz más grave que las mujeres.

Ejercicios

Elegir una hipótesis de cada grupo y buscar evidencia empírica a favor/en contra de cada una.

- Correlaciones:
 - Los hablantes de mayor edad suelen usar palabras de mayor longitud.
 - Los hablantes de mayor edad tienen menor dispersión (StDev) de nivel tonal.
- *t*-test de Student:
 - El habla espontánea tiene mayor dispersión (StDev) de intensidad que el habla leída.
 - Las mujeres hablan más lentamente que los hombres.
- Paired *t*-test:
 - El habla espontánea es más lenta que el habla leída.
 - El habla espontánea tiene un tono de voz más grave que el habla leída.

Aclaraciones:

La **tasa del habla** debe estimarse en cantidad de fonos por segundo, dentro de los segmentos hablados. Es decir, los silencios no deben tenerse en cuenta en este cómputo.

En los archivos *.ipu, los **silencios** se marcan con el símbolo “#”.