

ACÀMICA

¡Bienvenidos/as a Data Science!



Agenda

¿Cómo anduvieron?

Repaso: Outliers

¿Sabías que...?

Actividad: Dudas comunitarias

ENCUESTA

Break

Recursos: ¿Cómo dar feedback?

DEMO: Proyectos 1 y 2

Retro



¿Cómo anduvieron?



Repaso: Outliers



¿Qué es un
Outlier?
¿Por qué
ocurren?

OUTLIER = valor atípico que difiere significativamente del resto de las observaciones.

¿Por qué difiere?

- Error de medición del instrumento.
- Error al introducir un dato.
- Estamos trabajando con muestras/poblaciones que no son tan homogéneas como creíamos.



¿Qué es un
Outlier?
¿Por qué
ocurren?



¡Muchas veces los **OUTLIERS** son una manifestación del proceso que estamos estudiando!

Ejemplos:

- Transacción fraudulenta con una tarjeta de crédito.
- Persona enferma en un conjunto de personas sanas.
- Mayor incidencia de una enfermedad en una ciudad.
¿Esperable o outlier?



SUGERENCIA

Siempre es importante
pensar por qué hay un
outlier en nuestro dataset

Tipos de valores atípicos

univariante

Se desvía de los valores típicos de un feature (columna)

multivariante

Se desvía de los valores típicos que hay en la relación de dos o más columnas

Los valores atípicos suelen *confundir* la estadística que hacemos sobre los datos, ya que nos indican que no estamos trabajando con poblaciones homogéneas (que queremos detectar para remover).

A veces, **detectar outliers** es el objetivo de nuestro estudio.

¿Se les ocurre algún ejemplo?

¿Cómo detectar outliers?

Muchas veces no existe una manera *obvia* de detectar outliers, y, en general, ¡depende del problema!

Algunas técnicas son



- Visualización: Boxplots
- Por rango intercuartílico (Interquartile Range)
- Regla de las tres sigmas
- ¡Y más!

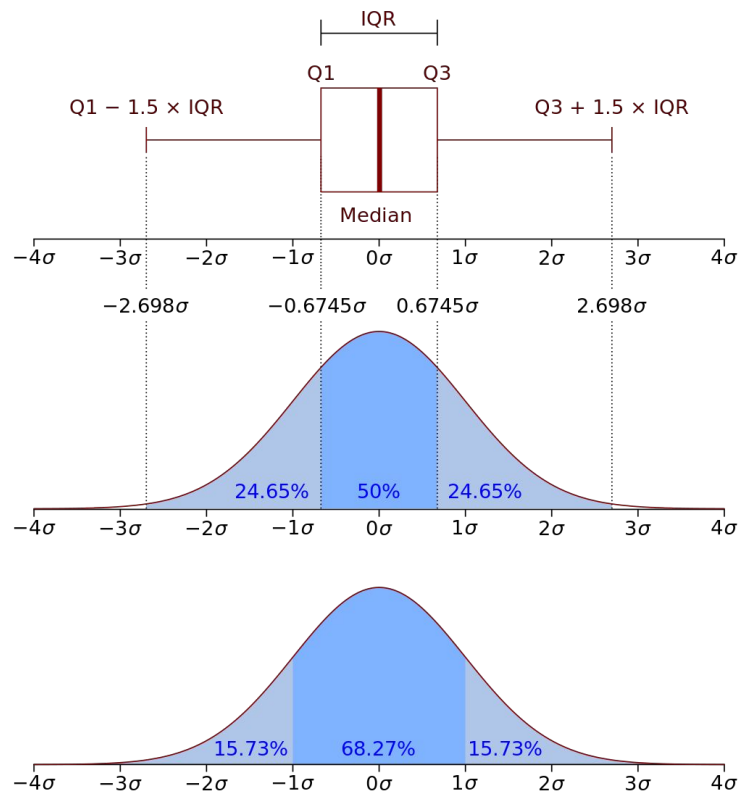
Boxplot

Rango intercuartílico

Regla de las tres sigmas

El **diagrama de cajas** es una forma de visualizar un conjunto de valores.

Muchas veces resulta más **informativa** que simplemente dibujar un punto por cada valor, ya que nos permite tener una idea de como es la distribución subyacente.

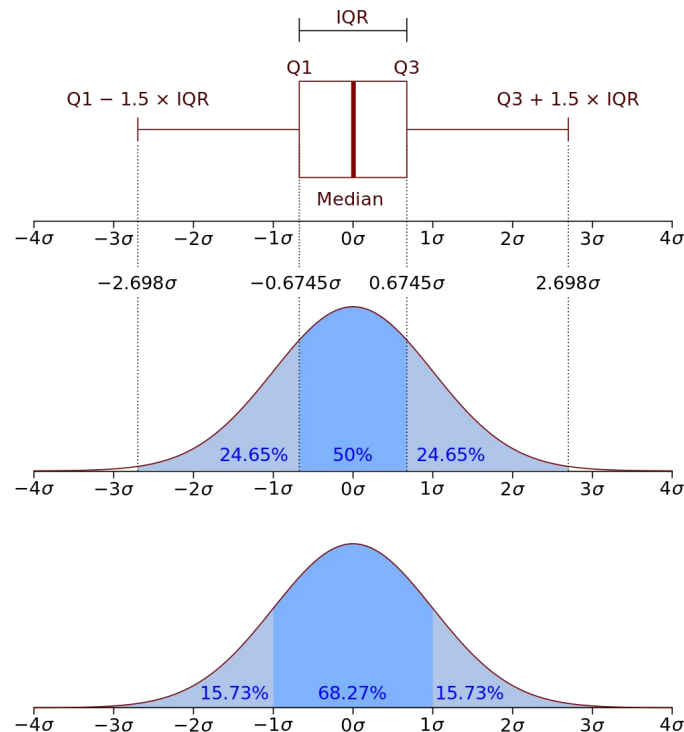


Elegimos un valor mínimo y un valor máximo para los valores “permitidos”.

Marcamos como outliers aquellos valores que estén por debajo del mínimo o por arriba del máximo.

¿Cómo elegimos el mínimo y el máximo?

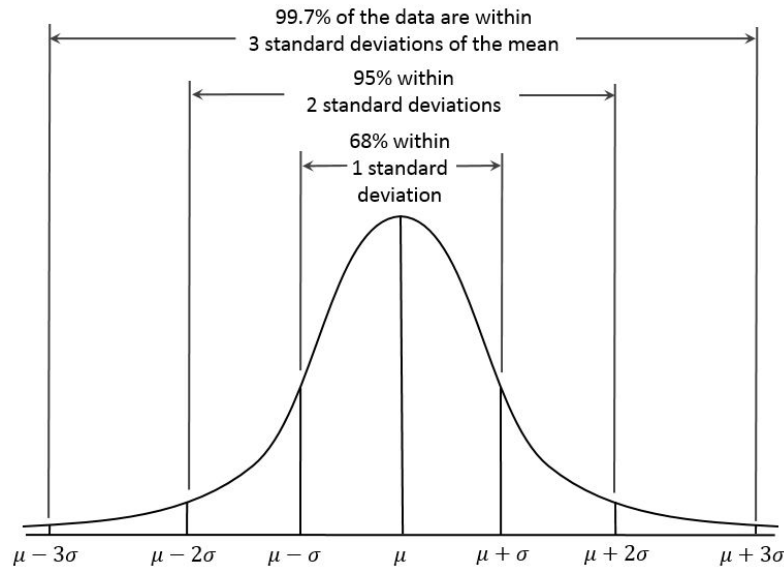
- A veces, es la variable la que nos lo indica. Por ejemplo, la asistencia a un curso no puede ser menor que cero o mayor al número de alumnos que tiene el curso.
- Un criterio estandarizado es usar
mínimo = $Q1 - 1.5 \times IQR$
máximo = $Q3 + 1.5 \times IQR$



¿Y si en lugar de usar los cuartiles
usamos las desviaciones estándar?

mínimo = valor medio - 3 x SD

máximo = valor medio + 3 x SD



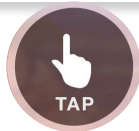
¿Sabías cuál es el problema de las ciudades chicas?



Educación

Pruebas Aprender: en Provincia, los mejores rendimientos los tienen los alumnos de las ciudades chicas y del interior

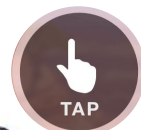
Clarín accedió a los resultados desagregados por municipios. Los distritos del sur y sureste de la Provincia están al tope de la lista, tanto en Lengua como en Matemática. Los más pobres del GBA quedaron relegados.



Santa Fe

Clarín en San Jorge, el pueblo donde los suicidios triplican el promedio de la Argentina

El drama afecta sobre todo a jóvenes y adolescentes. Lo atribuyen a varios factores y ya declararon la emergencia social.

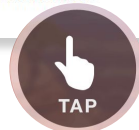


SOCIEDAD

28/04/2015 AGROTÓXICOS

Un pueblo de Entre Ríos en alerta: casi la mitad de su población muere por cáncer

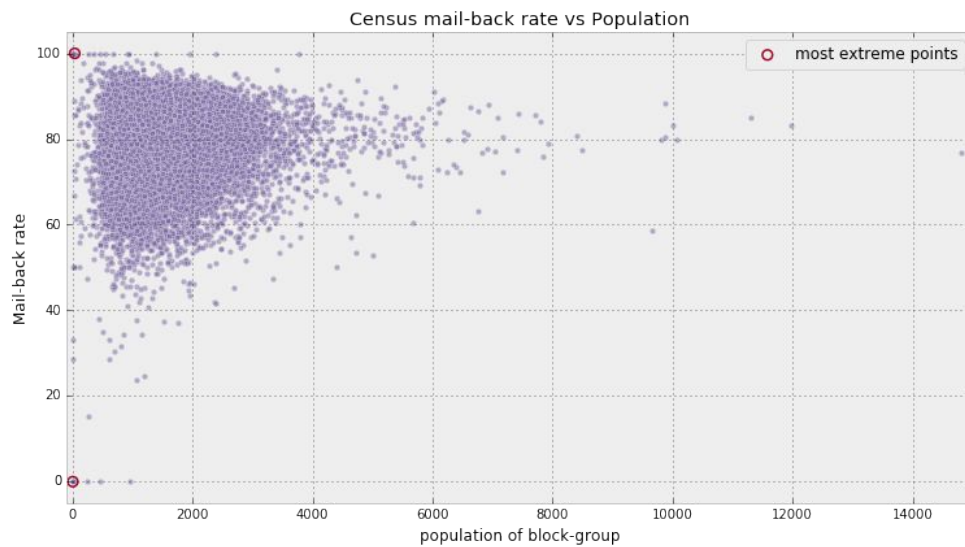
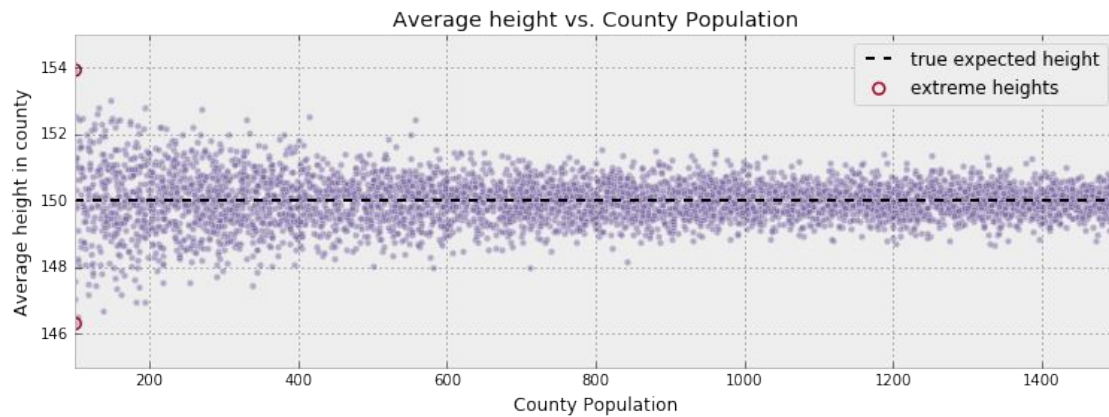
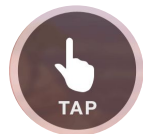
La fuerte movilización de habitantes que denuncian que casi la mitad de su población muere por cánceres generados presumiblemente por los agrotóxicos, motivó que el municipio entrerriano de San Salvador convocara a especialistas de las universidades de Rosario y de La Plata para realizar un estudio epidemiológico-ambiental, cuya primera etapa se cumplió la semana pasada, con científicos encuestando vecinos casa por casa y tomando muestras de aire, tierra y agua.



En los ejemplos que veremos a continuación, cabe preguntarnos...

¿Estamos en presencia de
fenómenos “reales” o se
tratan de desviaciones
estadísticas esperables?

Ejemplo 1

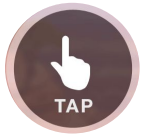


Ejemplo 2

The Most Dangerous Equation

*Ignorance of how sample size affects statistical variation
has created havoc for nearly a millennium*

Howard Wainer



¹ Este artículo es muy interesante, pero uno de los ejemplos es discutible.

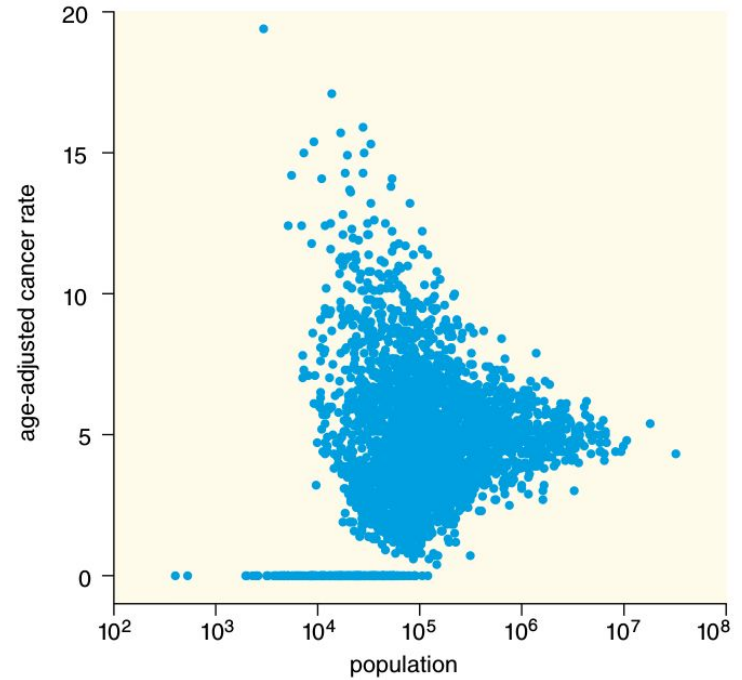
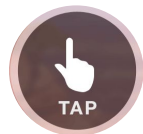


Figure 3. When age-adjusted kidney-cancer rates in U.S. counties are plotted against the log of county population, the reduction of variation with population becomes obvious. This is the typical triangle-shaped bivariate distribution.

Ejemplo 3



The most dangerous hospital or the most dangerous equation?

[Yu-Kang Tu](#)  & [Mark S Gilthorpe](#)

[BMC Health Services Research](#) 7, Article number: 185 (2007) | [Download Citation](#) ↓

5556 Accesses | 8 Citations | 2 Altmetric | [Metrics](#) >>

Results

A close examination of the information reveals a pattern which is consistent with a statistical phenomenon, discovered by the French mathematician de Moivre nearly 300 years ago, described in every introductory statistics textbook: namely that variation in performance indicators is expected to be greater in small Trusts and smaller in large Trusts. From a statistical viewpoint, the number of deaths in a hospital is not in proportion to the size of the hospital, but is proportional to the square root of its size. Therefore, it is not surprising to note that small hospitals are more likely to occur at the top and the bottom of league tables, whilst mortality rates are independent of hospital sizes.

Conclusion

This statistical phenomenon needs to be taken into account in the comparison of hospital Trusts performance, especially with regard to policy decisions.

Entonces...

Cuando tratamos con **muestras pequeñas**, ¡la **varianza es mucho mayor**!

Entonces hay que tener mucho cuidado al afirmar que hay una mayor (o menor) incidencia de un efecto en alguna de esas muestras *por encima de lo esperado*.

En general, para poder decir que existe un fenómeno de este tipo, vamos a tener que:

**Cuantificar que la desviación encontrada está por encima de lo esperado.
La estadística tiene herramientas que nos ayudan en esta tarea.**

Y es muy útil, además, una relación causal que pueda explicar el fenómeno que estamos viendo.

Esto dependerá de las características específicas de cada problemática.

En general, son problemáticas multicausales y existe un riesgo de caer en explicaciones simplistas o, directamente, erróneas.

Entrega 2: Transformación



Entrega 2: Transformación 75

Utilizá tu conocimiento del dominio para crear nuevos features

● Beginner by  Francisco Dorr

Actividad:

Dudas comunitarias



Encuesta

¡Queremos escucharte!





ENCUESTA



Para obtener el certificado, debes tener aprobados todos los proyectos de la carrera y haber completado todas las encuestas de cierre de bloque.

A close-up photograph of a white ceramic cup filled with a latte. The surface of the milk is decorated with intricate latte art, featuring a central heart shape surrounded by concentric, wavy lines. The cup is placed on a matching white saucer. In the background, a white napkin and a silver spoon are visible, though they are out of focus. The overall lighting is soft and even, highlighting the textures of the coffee and the smooth surface of the cup.

¡BREAK!

Ph. Credit: Drew Coffmann



Repaso:

¿Cómo dar feedback?



Pautas para esta sesión de feedback



Pautas para esta sesión de feedback

*"Me parece un efectivo que
hayas hecho xyz (**aspecto que
se está observando**)
porque (**resuelve el problema
xyz, aporta xyz, permite xyz**)"*



Pautas para esta sesión de feedback

*“Me parece un efectivo que
hayas hecho xyz (**aspecto que
se está observando**)
porque (**resuelve el problema
xyz, aporta xyz, permite xyz**)”*

ASPECTOS
POSITIVOS



ASPECTOS
A MEJORAR



*“Me parece que xyz (**aspecto
que se está observando**) NO es
efectivo porque (**NO resuelve el
problema xyz, NO aporta xyz,
NO permite xyz**)”*



DUDAS



IDEAS DE
MEJORA

Pautas para esta sesión de feedback

*"Me parece un efectivo que
hayas hecho xyz (**aspecto que
se está observando**)
porque (**resuelve el problema
xyz, aporta xyz, permite xyz**)"*

ASPECTOS
POSITIVOS



ASPECTOS
A MEJORAR



*"Me parece que xyz (**aspecto
que se está observando**) NO es
efectivo porque (**NO resuelve el
problema xyz, NO aporta xyz,
NO permite xyz**)"*

*"¿Por qué decidiste hacer xyz?"
"¿Probaste con xyz?"*

DUDAS



IDEAS DE
MEJORA



Pautas para esta sesión de feedback

*"Me parece un efectivo que
hayas hecho xyz (**aspecto que
se está observando**)
porque (**resuelve el problema
xyz, aporta xyz, permite xyz**)"*

ASPECTOS
POSITIVOS



ASPECTOS
A MEJORAR



*"Me parece que xyz (**aspecto
que se está observando**) NO es
efectivo porque (**NO resuelve el
problema xyz, NO aporta xyz,
NO permite xyz**)"*

*"¿Por qué decidiste hacer xyz?"
"¿Probaste con xyz?"*

DUDAS



IDEAS DE
MEJORA



*"Puedes probar usando la
herramienta xyz o la
metodología wmb."*

*"En lugar de x haría y para
(**razón para hacerlo**)."*

*"Puedes probar con xyz. Algo
similar se hizo (dar un
ejemplo)."*

DEMO:

Proyecto 01

(Entregas 01 y 02)





¡FELICIDADES!

RETRO:

mantener, mejorar, probar

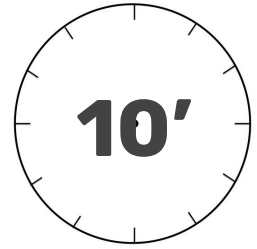


Retro

Al final de todo ciclo de desarrollo ágil, existe una ceremonia llamada “Retro” (por retrospectiva) que nos permite construir feedback constructivo entre el grupo para poder aplicar puntos de acción en la siguiente iteración.



Paso 1



Escribir en post its/papel de manera individual pensamientos sobre lo que crees que hay que mantener, mejorar y probar en relación a:

- dinámicas de clase
- trabajo en los proyectos
- dedicación y trabajo propios
- cualquier otra cosa que consideren decir!



MANTENER

¿Qué es lo que salió bien hasta ahora y puede servir para más adelante?

yo



MEJORAR

¿Qué es lo que no salió tan bien hasta ahora y que hay que hacer de otra manera?

**como
clase**



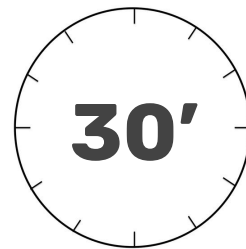
PROBAR

¿Qué es lo que no hice /se hizo hasta ahora pero quizás puede servir intentar?





Paso 2



¡Momento de compartir!

La idea es que todos puedan participar y decir lo que escribieron. Para hacerlo de manera ordenada, haremos de a uno por vez.

Si escuchas un pensamiento similar al tuyo, ¡súmate instantáneamente a él y comparte el que tu escribiste.

Si escuchas algo que no incluiste pero te parece valioso, demuéstalo.

Para la próxima

1. Terminar la Entrega 02.
2. ¡Arrancamos con Machine Learning! Ver los videos “Machine Learning: Qué es Machine Learning”
3. Completar notebooks atrasados y, si lo desean, seguir explorando el dataset que eligieron.

ACÀMICA