03_Estadistica_III_ConceptosAvanzados

January 7, 2026

Contenido

1 ESTADÍSTICA III PARA DATA SCIENCE: CONCEPTOS AVANZADOS

1.1 CONCEPTOS AVANZADOS

En esta sección vamos a ver conceptos avanzados de dos tipos:

- conceptos más avanzados que estaremos usando sobre todo cuando hagamos modelización estadística
- errores de interpretación en el día a día de las conclusiones que se sacan en una empresa, pero que no son correctas.

1.1.1 Valor teórico y error de medida

El valor teórico es una combinación lineal de variables con ponderaciones estimadas empíricamente.

Es decir, cuando hagamos modelos predictivos que vienen del campo de la estadística, como regresiones, lo que estará prediciendo el modelo es el valor teórico.

Pero el valor teórico se puede descomponer en dos componentes: la parte "verdadera"de la variable objetivo que se puede predecir directamente a través de las predictoras, y un error.

Ese error se llama **error de medida** y es lo que siempre estaremos intentando minimizar.

A nivel operativo ese error tendrá dos fuentes principales.

Una es la falta de datos.

Y la otra es la falta de adecuación del algoritmo que estemos usando.

Por tanto en data science siempre mejoraremos:

- Incorporando nuevos y/o mejores datos
- Utilizando algoritmos más apropiados al problema

Y a nivel conceptual ese error puede tener dos fuentes, que son imporantes para detectar malas prácticas en la empresa:

- La validez: que estemos midiendo lo que en realidad queremos medir
- La fiabilidad: que lo estemos midiendo con los instrumentos adecuados

Por ejemplo un error de validez en la empresa es cuando las agencias intentan convencer a los clientes de que invertir en campañas de "likes" mejorará sus resultados comerciales.

Y un ejemplo de error de fiabilidad es intentar medir la importancia de cada canal en nuestro marketing mediante informes basados en atribución "last click".

1.1.2 Supuestos estadísticos

Las técnicas que utilizaremos en la parte de modelización vienen de dos campos:

- Machine Learning: como árboles de decisión, random forest, redes neuronales, …
- De la estadística: como regresión múltiple, regresión logística

Estas últimas fueron creadas bajo una serie de supuestos, que idealmente deberían cumplirse para que se pueda utilizar la técnica.

Es más, los supuestos deberían cumplirse tanto a nivel individual de cada variable como en el valor teórico combinado.

Pero en la práctica es muy raro que estos supuestos se cumplan, y realmente las técnicas han demostrado ser robustas a las violaciones de los mismos.

En su momento aprenderemos la forma de evaluar los modelos en la práctica para ver si, aún sin cumplir totalmente los supuestos, son modelos útiles o no.

Pero es importante conocerlos al menos a nivel conceptual.

Ello nos llevará a entender mucho mejor este tipo de técnicas y a ser capaz de mejorar su capacidad predictiva.

Los supuestos más importantes son:

- Normalidad
- Heterocedasticidad de varianzas
- Linealidad

- Multicolinealidad

Normalidad Ya conocemos lo que es una distribución normal.

Si la variación con respecto a una normal es suficientemente grande, entonces todos los test estadísticos resultantes (que se basarán en estadísticos como t o F) no serán válidos.

Medir la normalidad multivariante es complicado, por lo que la aproximación práctica suele ser medir y corregir la normalidad univariante de todas las variables que formarán el valor teórico.

Formas de identificarla:

- La normalidad se puede evaluar a nivel gráfico con gráficos como el histograma o el Q-Q.

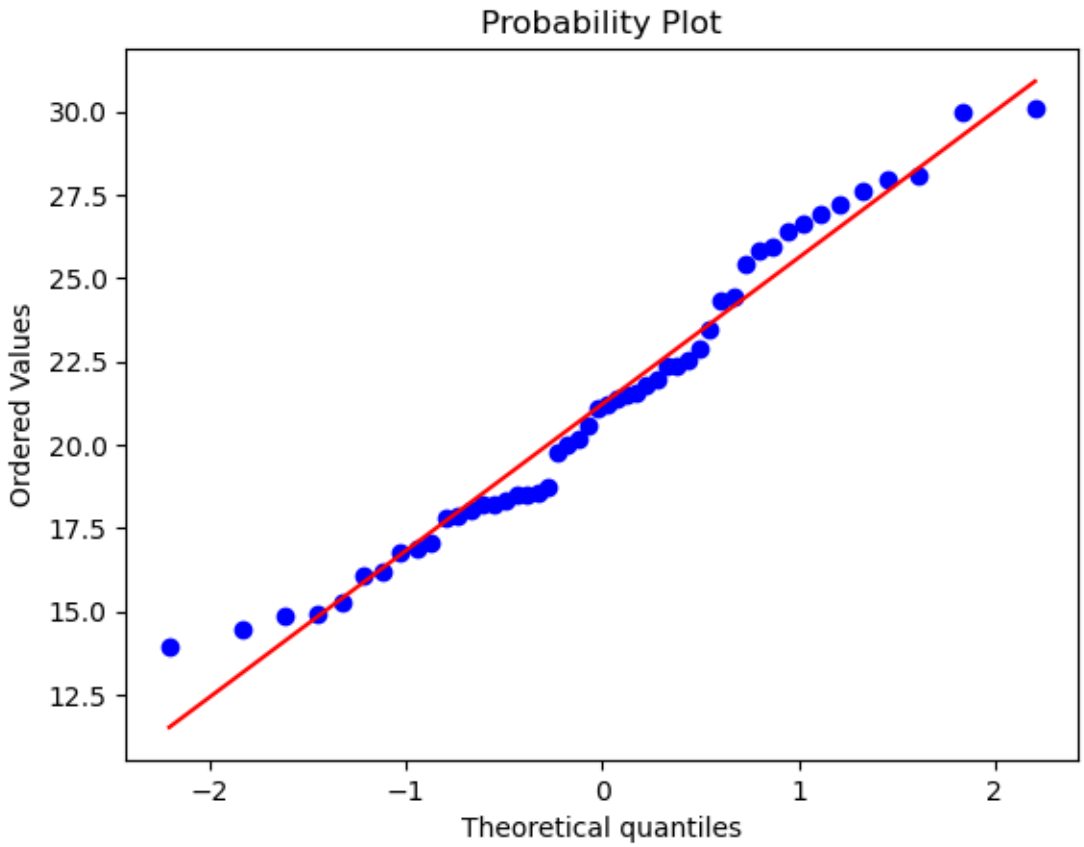
Solución:

- Transformaciones de la variable para hacerla más normal: inversa, cuadrado, raiz cuadrada, logaritmo, ...

[2]:
```
#Ejemplo de un qqplot

import matplotlib.pyplot
import numpy as np
import scipy.stats

ejemplo = np.random.normal(loc = 20, scale = 5, size=50)

scipy.stats.probplot(ejemplo, dist="norm", plot=matplotlib.pyplot)
```

[2]:
```
((array([-2.20385432, -1.83293478, -1.61402323, -1.45296849, -1.32267759,
         -1.21163342, -1.113805  , -1.02561527, -0.94475674, -0.86964726,
         -0.79915021, -0.73241807, -0.66879925, -0.6077796 , -0.54894415,
         -0.49195112, -0.43651377, -0.38238727, -0.32935914, -0.27724191,
         -0.2258675 , -0.17508277, -0.12474591, -0.07472335, -0.02488719,
          0.02488719,  0.07472335,  0.12474591,  0.17508277,  0.2258675 ,
          0.27724191,  0.32935914,  0.38238727,  0.43651377,  0.49195112,
          0.54894415,  0.6077796 ,  0.66879925,  0.73241807,  0.79915021,
          0.86964726,  0.94475674,  1.02561527,  1.113805  ,  1.21163342,
          1.32267759,  1.45296849,  1.61402323,  1.83293478,  2.20385432]),
  array([13.93031797, 14.45486057, 14.87082826, 14.89778484, 15.25864167,
         16.04186619, 16.20234493, 16.78798232, 16.86271757, 17.04887429,
         17.82217731, 17.85138853, 18.01134503, 18.20300066, 18.23051415,
         18.34150515, 18.46960765, 18.48918722, 18.53714566, 18.71213582,
         19.76921453, 19.99225015, 20.18562777, 20.57069155, 21.07874659,
         21.21483261, 21.38077968, 21.50318295, 21.55409907, 21.76610722,
         21.97304366, 22.35971533, 22.37881118, 22.55301975, 22.8747798 ,
         23.46514211, 24.32646361, 24.43486588, 25.40937145, 25.83441726,
         25.92644189, 26.40846189, 26.62463148, 26.93078583, 27.22948419,
         27.63509348, 27.97786149, 28.07943407, 29.9706434 , 30.10999383])),
 (np.float64(4.398892280283699),
```

```
np.float64(21.21084438985098),
np.float64(0.9866018374926581)))
```



Heterocedasticidad de varianzas Significa que la varianza de la variable objetivo no es constante en el recorrido de la variable predictora.

Es decir, que para unos valores la predicción será más precisa que para otros.

Formas de identificarla:

- Para variables contínuas: con diagramas de dispersión
- Analizando el gráfico de los residuos (diferencias entre valor predicho y valor real)

Solución:

- En la mayoría de casos es causada por la no normalidad, por lo que corrigiendo la normalidad se corrige también la heterocedasticidad.

Linealidad Se refiere a que exista una relación lineal de cada variable predictora con la target.

Es un supuesto para todas las técnicas que se basen de una u otra forma en la correlación, como la regresión múltiple o la regresión logística.

Formas de identificarla:

- Hacer la matriz de correlaciones de cada predictora con la target
- Gráficos de dispersión de cada predictora con la target
- Análisis de residuos del modelo. Cualquier pauta no lineal visible será la que las variables no han podido explicar linealmente

Solución:

- Linealizar las relaciones mediante transformación de las variables originales
- Usar algoritmos no lineales

Multicolinealidad Se refiere a que exista correlación entre las variables predictoras.

Los modelos que usamos normalmente (aditivos) asumen independencia entre las variables predictoras.

Si eso no se cumple pasará que:

- Podemos estar sobreponderando conceptos
- Podemos causar efectos extraños y variaciones en los modelos: coeficientes desproporcionados, signos invertidos, o incluso no convergencia

Formas de identificarla:

- Hacer la matriz de correlaciones entre las variables predictoras
- Identificar durante el desarrollo del modelo variables que apriori deberían predecir pero salen como no predictoras (por la correlación parcial aportada por otras variables)

Solución:

- No meter variables correlacionadas
- Aplicar reducción de variables como Componentes Principales (poco uso en la realidad)

1.1.3 Bootstrapping

Si recuerdas ya introdujimos este concepto cuando hablamos del muestreo.

Y es que de hecho se traduce al español como re-muestreo.

Ya vimos cómo el remuestreo había conseguido demostrar el teorema del límite central, y a partir de ahí todo lo que vimos en la parte inferencial.

Pero tiene más propiedades. Y una de ellas es la capacidad de generalización.

Como veremos en su momento el mayor enemigo que vamos a tener es el sobre ajuste.

Pues bien, hay algoritmos que se basan precisamente en bootstrapping para conseguir evitar ese sobreajuste y ser capaz de predecir mucho mejor ante datos que no han visto nunca.

Lo usan por ejemplo el bagging, que viene de bootstrap aggregating, y básicamente consiste en remuestrear casos y / o variables, hacer un modelo sobre cada una de esas muestras y luego agregarlos para hacer la predicción final.

Por ejemplo Random Forest, que veremos en la parte de modelización, se basa en este principio y es de los algoritmos más estables.

1.1.4 Penetración vs distribución

No es estrictamente un concepto estadístico, pero podemos meterlo dentro de los análisis que más se confunden.

Este efecto está implicado en falacias como:

"Según la DGT el alcohol está implicado en el 30% de los accidentes mortales".

Luego el alcohol no está implicado en el 70% de los accidentes mortales.

Por tanto podemos concluir que es más peligroso conducir sobrio que borracho.

O su equivalente muy frecuente en el mundo de la empresa:

"El 70% de nuestras compras las hacen clientes varones entre 30 y 45 años, por tanto son los más compradores"

No será cierto si ese perfil fuera el 80% del total de clientes por ejemplo.

En general hay que diferenciar muy bien estos dos conceptos:

- Penetración: qué porcentaje de una base determinada presenta la característica a analizar
- Distribución: qué porcentaje representa un sobconjunto sobre el total

El mejor truco para diferenciarlo es que al sumar penetraciones no tiene por qué sumar 100%.

Pero al sumar distribuciones sí tiene que sumar 100%.

Por ejemplo, decir que el 70% de nuestros clientes son mujeres es un análisis de distribución, y tiene que sumar 100% con el 30% restante que son hombres.

Sin embargo decir que el producto A lo compra el 70% de las mujeres y el 50% de los hombres es un análisis de penetración, y no tiene por qué sumar 100%.

Estos dos son conceptos absolutamente claves cuando estemos haciendo análisis de perfilado, insights o segmentaciones.

1.1.5 Absoluto Vs relativo

3 fuentes de error:

- dar el absoluto sin dar el total
- dar el porcentaje sin dar el término de comparación absoluto
- no tener en cuenta las escalas de los gráficos

Con esto muchas veces se intenta hacer trampas, usando uno u otro según convenga al mensaje que se quiere mandar.

Por ejemplo dar un dato en absoluto, sin dar el total, hará que su efecto parezca mayor o menor.

Decir que este fin de semana han muerto 50 personas en carretera parece mucho. Pero si decimos que ha sido salida de Semana Santa, con 15.000.000 millones de desplazamientos el dato queda matizado.

O al revés, dar el dato en relativo a sabiendas de que la base original es muy pequeña y por tanto parecerá un efecto mayor.

Por ejemplo si decimos que esta semana hemos vendido un 300% más que la anterior parece mucho, pero si la anterior sólo vendimos un producto el dato ya no impresiona tanto.

O jugar con los ejes para que en las comparaciones se perciba un efecto diferente al real, como en este gráfico de RTVE.

[3]:
```
from IPython import display
display.Image("../98_Media/falsear con grafico.png")
```

[3]:

Solución:

- No considerar datos en porcentaje que tengan una base menor a 50 casos
- Siempre poner en contexto ambas dimensiones
- Cuidado con los ejes y las posiciones relativas o absolutas

1.2 REPASO DE LO APRENDIDO

- Descriptiva vs inferencial

- Población y muestra
- Escalas de medida y tipos de variables
- Principales estadísticos para analizar, y cuando usar cada uno
- Principales gráficos para analizar, y cuando usar cada uno
- Chi-cuadrado y correlaciones, y cuando usar cada uno
- Principales distribuciones
- Teorema del límite central y todo lo que supone
- Cómo calcular intervalos de confianza
- Cómo calcular el tamaño de la muestra
- Qué son el Alpha y el Pvalor
- Por qué se hace lo de comparar el pvalor con el Alpha
- Qué son las hipótesis nula y alternativa y cómo se contrastan
- Cómo hacer un contraste de hipótesis de medias en la población
- Cómo hacer un contraste de hipótesis de proporciones en la población
- Cómo hacer un contraste de hipótesis de medias entre dos muestras
- Cómo hacer un contraste de hipótesis de proporcinoes entre dos muestras
- Los principales supuestos estadísticos de los modelos, cómo se identifican y cómo se corrigen
- Fallos más comunes de interpretación y errores a evitar: correlación vs causalidad, penetración vs distribución, absluto vs relativo, etc