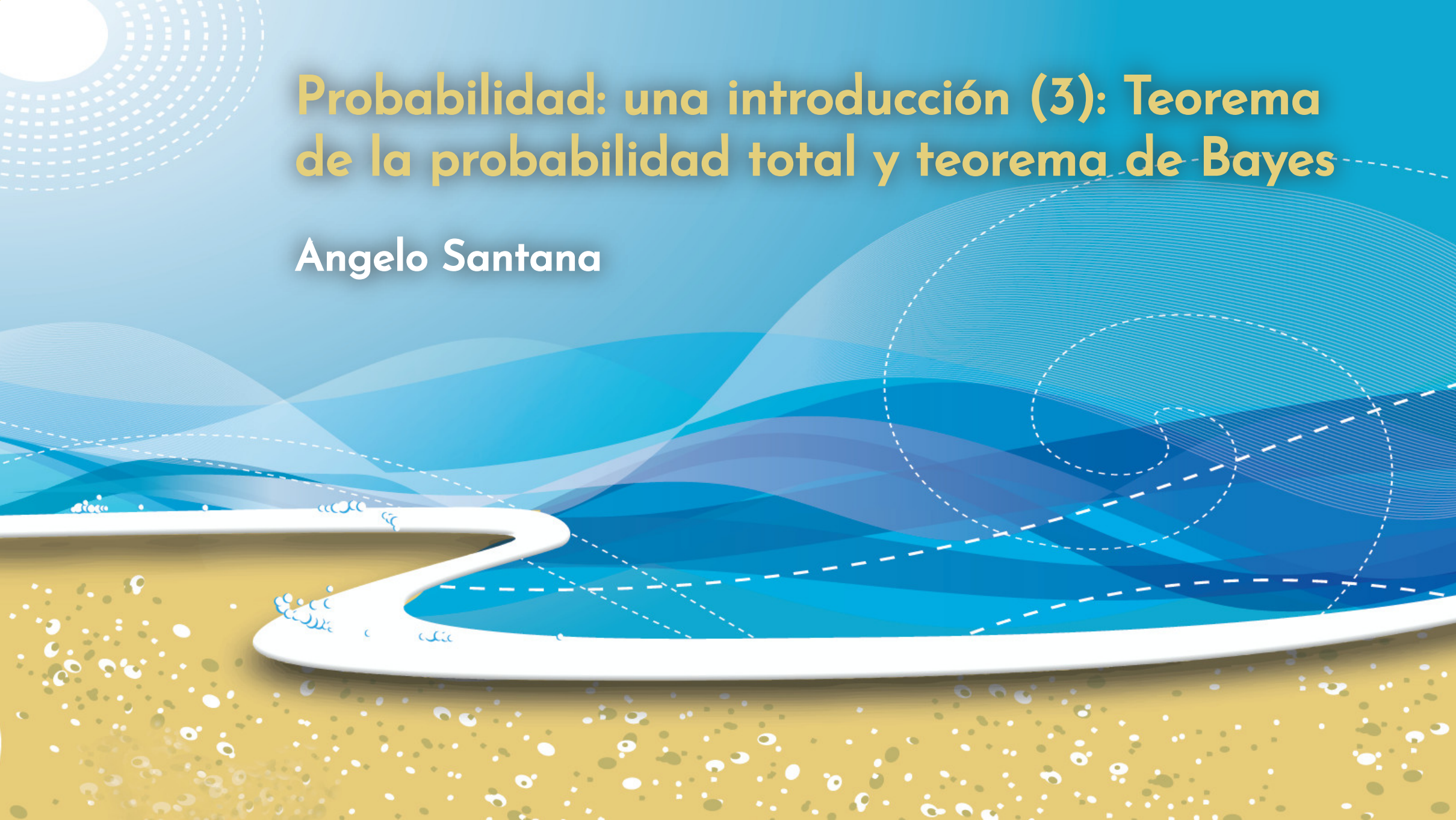
Probabilidad: una introducción (3): Teorema de la probabilidad total y teorema de Bayes

Angelo Santana



Ejemplo 3. Análisis de calidad de aguas de baño.

Normalmente el agua de las playas de uso público se somete de modo periódico a la realización de análisis para detectar la presencia de posibles contaminantes en niveles peligrosos para la salud humana.

La presencia del contaminante suele llevar aparejado el cierre de la playa, mientras que la no presencia del mismo significa que la playa se considera apta para el baño.

Ejemplo 3. Análisis de calidad de aguas de baño.

- En este contexto se consideran los siguientes sucesos:
  - **D**: el análisis da positivo, es decir, **detecta** la presencia del contaminante en el agua.
  - **C**: el agua está **contaminada**.
- **IMPORTANTE:** el análisis químico no es infalible: a veces dará positivo sin que haya contaminación, y a veces dará negativo aunque el agua se encuentre contaminada.

$$P(D|C) = p_V \qquad P\left(\overline{D}|C\right) = 1 - p_V$$

$$P\left(D|\overline{C}\right) = p_F \qquad P\left(\overline{D}|\overline{C}\right) = 1 - p_F$$

Ejemplo 3. Análisis de calidad de aguas de baño.

- Supongamos que la experiencia acumulada a lo largo del tiempo indica que la playa presenta contaminación una proporción $P(C)$ de los días del año; o dicho de otro modo, en un día elegido al azar, la probabilidad de que la playa se encuentre contaminada es $P(C)$.

- Nos planteamos las siguientes cuestiones:

  - ¿Cuál es la probabilidad de que el procedimiento de análisis dé positivo (detecte contaminación en la playa) en un día arbitrario? Es decir, ¿cuánto vale $P(D)$?

  - Si el análisis ha dado positivo, ¿cuál es la probabilidad de que realmente la playa esté contaminada ese día?. Es decir, ¿cuánto vale $P(C|D)$?

Teorema de la probabilidad total.

Sea $\{A_1, \dots A_n\}$ un *sistema completo de sucesos*, es decir, una colección de sucesos definidos sobre un espacio muestral Ω tales que:

- $A_1 \cup A_2 \cup \cdots \cup A_n = \Omega$
- $A_i \cap A_j = \phi \quad \forall i, j$

Sea B un suceso arbitrario de Ω. El **teorema de la probabilidad total** especifica que:

$$\Pr(B) = \sum_{i=1}^{n} \Pr(B|A_i) \Pr(A_i)$$

Demostración:

$$B = \Omega \cap B = (A_1 \cup A_2, \cup A_n) \cap B = (A_1 \cap B) \cup \cdots \cup (A_n \cap B)$$

Por tanto:

$$\Pr(B) = P((A_1 \cap B) \cup \cdots \cup (A_n \cap B))$$

y como los sucesos $(A_i \cap B)$ son incompatibles dos a dos, se tiene que:

$$\Pr(B) = P((A_1 \cap B) \cup \cdots \cup (A_n \cap B)) = \sum_{i=1}^{n} \Pr(A_i \cap B)$$

Por último, como $\Pr(A_i \cap B) = \Pr(B \mid A_i) \Pr(A_i)$, se sigue que:

$$\Pr(B) = \sum_{i=1}^{n} \Pr(B \mid A_i) \Pr(A_i)$$

Aplicación

¿Cuál es la probabilidad total de detectar contaminación en el agua?

C y $\overline{C}$ forman un sistema completo de sucesos, y por tanto:

$$P(D) = P(D|C)\,P(C) + P\left(D\middle|\overline{C}\right)P\left(\overline{C}\right)$$

Si, por ejemplo: $P(D|C) = 0.99$, $P\left(D\middle|\overline{C}\right) = 0.10$, $P(C) = 0.03$, entonces:

- $P\left(\overline{C}\right) = 1 - P(C) = 1 - 0.03 = 0.97$
- $P(D) = 0.99 \cdot 0.03 + 0.10 \cdot 0.97 = 0.0297 + 0.097 = 0.1267$
- ¡¡Nótese que en este ejemplo sólo hay contaminación un 3% de los días, pero el análisis la detecta en algo más de un 12%!! Esto significa que algo más de un 9% de las detecciones son *falsos positivos*.

Teorema de Bayes

Thomas Bayes en Wikipedia

Sea $\{A_1, \dots A_n\}$ un sistema completo de sucesos definido sobre un espacio muestral Ω, y sea B un suceso arbitrario de Ω. Entonces:

$$\Pr(A_i \mid B) = \frac{\Pr(B \mid A_i)\Pr(A_i)}{\Pr(B)} =$$

$$= \frac{\Pr(B \mid A_i)\Pr(A_i)}{\sum_{j=1}^{n} \Pr(B \mid A_j)\Pr(A_j)}$$

Demostración

De acuerdo con la definición de probabilidad condicionada:

$$\Pr(A_i \mid B) = \frac{P\left(A_i \cap B\right)}{P\left(B\right)}$$

- También de la definición de probabilidad condicionada se sique que el numerador puede expresarse como $P\left(A_i \cap B\right) = P\left(B \mid A_i\right) P\left(A_i\right)$
- De acuerdo con el teorema de la probabilidad total, el denominador es $\Pr\left(B\right) = \sum_{i=1}^{n} \Pr\left(B \mid A_i\right) \Pr\left(A_i\right)$

Sustituyendo numerador y denominador por estas expresiones, obtenemos el teorema de Bayes.

Aplicación

Si el análisis ha dado positivo, ¿cuál es la probabilidad de que realmente la playa esté contaminada ese día?. Es decir, ¿cuánto vale $P(C|D)$?

Aplicando el teorema de Bayes:

$$P(C \mid D) = \frac{P(D \mid C)\, P(C)}{P(D)} = \frac{0.99 \cdot 0.03}{0.1267} = 0.2344$$

Por tanto, si el análisis ha dado positivo, la probabilidad de que realmente haya contaminación ¡¡es solo 0.2344!!

¿Como mejorar el protocolo de detección de contaminantes en el agua de la playa?

- Mejorando el procedimiento analítico para que el análisis sea más preciso. **¡La mejora o el desarrollo de nuevos procedimientos es complicado!**
- Repitiendo el análisis. **¿Cuántas veces?**

Mejora del protocolo de detección de contaminantes

- Un protocolo de detección de contaminantes es una regla que especifica, a partir de la información disponible, cuándo decidimos que la playa está contaminada y cuando decidimos que no.
- Denotemos por D^+ la decisión positiva (la playa está contaminada) y D^- la decisión negativa (la playa está limpia).
- El protocolo debe diseñarse de forma que los dos siguientes tipos de error sean pequeños:

  1. $P\left(C \,|D^-\right) \leq \alpha$
  2. $P\left(\overline{C} \,|D^+\right) \leq \beta$

- **¡¡Los dos tipos de error no tienen por qué tener la misma importancia!!**

Mejora del protocolo de detección de contaminantes

De acuerdo con el teorema de Bayes:

$$P\left(C\left|D^{-}\right.\right)=\frac{P\left(D^{-}\left|C\right.\right)P\left(C\right)}{P\left(D^{-}\right)}$$

$$P\left(\overline{C}\left|D^{+}\right.\right)=\frac{P\left(D^{+}\left|\overline{C}\right.\right)P\left(\overline{C}\right)}{P\left(D^{+}\right)}$$

siendo:

$$P\left(D^{+}\right)=P\left(D^{+}\left|C\right.\right)P\left(C\right)+P\left(D^{+}\left|\overline{C}\right.\right)P\left(\overline{C}\right)$$

$$P\left(D^{-}\right)=1-P\left(D^{+}\right)$$

Mejora del protocolo de detección de contaminantes

Si hacemos un único análisis y decidimos que la playa está contaminada cuando éste da positivo, ya hemos visto que en el caso particular de que $P(D|C) = 0.99$, $P\left(D\middle|\overline{C}\right) = 0.10$ y $P(C) = 0.03$, entonces:

- $P(D^+) = P(D) = 0.99 \cdot 0.03 + 0.10 \cdot 0.97 = 0.0297 + 0.097 = 0.1267$
- $P(D^-) = 1 - P(D) = 0.8733$
- $P(C|D^-) = \frac{P(D^-|C)P(C)}{P(D^-)} = \frac{0.01 \cdot 0.03}{0.8733} = 0.00034$
- $P\left(\overline{C}|D^+\right) = \frac{P\left(D^+\middle|\overline{C}\right)P\left(\overline{C}\right)}{P(D^+)} = \frac{0.10 \cdot 0.97}{0.1267} = 0.7656$

Como vemos, la probabilidad de que la playa esté contaminada cuando se decide que no lo está es muy baja (**0.00034**); a cambio hay una probabilidad

Mejora del protocolo de detección de contaminantes

- Supongamos ahora que decidimos tomar **5** muestras de agua. Llamamos N_D al número de muestras en las que se detecta contaminación, y adoptamos la siguiente regla de decisión:
  - Si $N_D \geq 3$ (es decir 3 ó más dan positivo) decidimos D^+ (hay contaminación).
  - Si $N_D < 3$ decidimos D^- (la playa está limpia).

En estas condiciones:

- Si la playa está contaminada: $N_D \approx B\left(n = 5, p = P\left(D \middle| C\right) = 0.99\right)$
- Si la playa está limpia: $N_D \approx B\left(n = 5, p = P\left(D \middle| \overline{C}\right) = 0.10\right)$

Mejora del protocolo de detección de contaminantes

Entonces:

- $P\left(D^{-}\,|C\right) = P\left(N_D < 3\,|C\right) = \sum_{k=0}^{2} P\left(N_D = k\,|C\right) =$

$$= \sum_{k=0}^{2} \binom{5}{k} 0.99^{k} \cdot 0.01^{5-k} = \text{sum(dbinom(0 : 2, 5, 0.99))} = 9.85 \cdot 10^{-6}$$

- $P\left(D^{+}\,\middle|\,\overline{C}\right) = P\left(N_D \geq 3\,\middle|\,\overline{C}\right) = \sum_{k=3}^{5} P\left(N_D = k\,\middle|\,\overline{C}\right) =$

$$= \sum_{k=3}^{5} \binom{5}{k} 0.1^{k} \cdot 0.9^{5-k} = \text{sum(dbinom(3 : 5, 5, 0.10))} = 0.00856$$

Mejora del protocolo de detección de contaminantes

Además:

- $P\left(D^{-}\,|C\right) = 9.85 \cdot 10^{-6} \Rightarrow P\left(D^{+}\,|C\right) = 1 - 9.85 \cdot 10^{-6} = 0.9999901$
- $P\left(D^{+}\,\middle|\,\overline{C}\right) = 0.00856 \Rightarrow P\left(D^{-}\,\middle|\,\overline{C}\right) = 1 - 0.00856 = 0.99144$

y por tanto:

- $P\left(D^{+}\right) = P\left(D^{+}\,|C\right) P\left(C\right) + P\left(D^{+}\,\middle|\,\overline{C}\right) P\left(\overline{C}\right) =$

$$= 0.9999901 \cdot 0.03 + 0.00856 \cdot 0.97 = 0.0383$$

- $P\left(D^{-}\right) = 1 - P\left(D^{+}\right) = 1 - 0.0383 = 0.9617$

Mejora del protocolo de detección de contaminantes

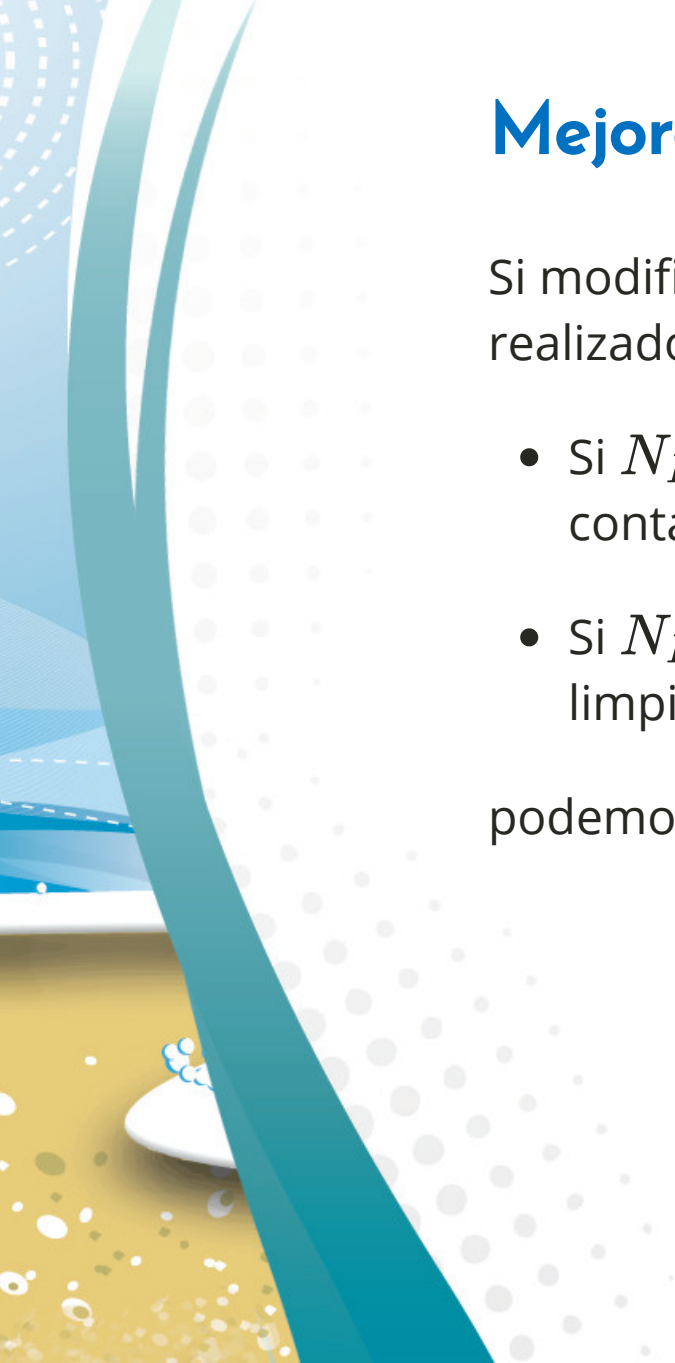
Finalmente:

$$P\left(C\left|D^{-}\right.\right)=\frac{P\left(D^{-}\left|C\right.\right)P\left(C\right)}{P\left(D^{-}\right)}=\frac{9.85\cdot10^{-6}\cdot0.03}{0.9617}=3.07\cdot10^{-7}$$

$$P\left(\overline{C}\left|D^{+}\right.\right)=\frac{P\left(D^{+}\left|\overline{C}\right.\right)P\left(\overline{C}\right)}{P\left(D^{+}\right)}=\frac{0.00856\cdot0.97}{0.0383}=0.2168$$

Vemos que con esta regla de decisión disminuyen las probabilidades de los dos tipos de error:

- El primero disminuye de 0.00034 a $3.07\cdot10^{-7}$
- El segundo disminuye de 0.7656 a 0.2168



Mejora del protocolo de detección de contaminantes

Si modificamos ligeramente la regla de decisión, de tal forma que una vez realizados los cinco análisis:

- Si $N_D > 3$ (es decir más de 3 dan positivo) decidimos D^+ (hay contaminación).
- Si $N_D \leq 3$ (es decir dan positivo 3 o menos) decidimos D^- (la playa está limpia).

podemos repetir los cálculos anteriores, y ahora obtenemos:

Mejora del protocolo de detección de contaminantes

- $P\left(D^{-}\,|C\right) = P\left(N_D \leq 3\,|C\right) = \sum_{k=0}^{3} P\left(N_D = k\,|C\right) =$

$$= \sum_{k=0}^{3} \binom{5}{k} 0.99^{k} \cdot 0.01^{5-k} = \text{sum(dbinom(0 : 3, 5, 0.99))} = 0.00098$$

- $P\left(D^{+}\,\middle|\overline{C}\right) = P\left(N_D > 3\,\middle|\overline{C}\right) = \sum_{k=4}^{5} P\left(N_D = k\,\middle|\overline{C}\right) =$

$$= \sum_{k=4}^{5} \binom{5}{k} 0.1^{k} \cdot 0.9^{5-k} = \text{sum(dbinom(4 : 5, 5, 0.10))} = 0.00046$$

Mejora del protocolo de detección de contaminantes

Entonces:

- $P\left(D^{-}\,|C\right) = 0.00098 \Rightarrow P\left(D^{+}\,|C\right) = 1 - 0.00098 = 0.99902$
- $P\left(D^{+}\,\middle|\,\overline{C}\right) = 0.00046 \Rightarrow P\left(D^{-}\,\middle|\,\overline{C}\right) = 1 - 0.00046 = 0.99954$

y por tanto:

- $P\left(D^{+}\right) = P\left(D^{+}\,|C\right) P\left(C\right) + P\left(D^{+}\,\middle|\,\overline{C}\right) P\left(\overline{C}\right) =$

$$= 0.99902 \cdot 0.03 + 0.00046 \cdot 0.97 = 0.0304$$

- $P\left(D^{-}\right) = 1 - P\left(D^{+}\right) = 1 - 0.0304 = 0.9696$

Mejora del protocolo de detección de contaminantes

Finalmente:

$$P\left(C\,|D^-\right) = \frac{P\left(D^-\,|C\right)P\left(C\right)}{P\left(D^-\right)} = \frac{0.00098 \cdot 0.03}{0.9696} = 0.0000303$$

$$P\left(\overline{C}\,|D^+\right) = \frac{P\left(D^+\,|\overline{C}\right)P\left(\overline{C}\right)}{P\left(D^+\right)} = \frac{0.00046 \cdot 0.97}{0.0304} = 0.0147$$

Con esta regla de decisión la probabilidad de que la playa esté contaminada cuando hemos decidido que está limpia es 0.0000303, mientras que la probabilidad de que esté limpia cuando hemos decidido que está contaminada es ya sólo de 0.0147.

Puede ser una regla de decisión aceptable si consideramos que el riesto que representan estas probabilidades es asumible.

Mejora del protocolo de detección de contaminantes

La siguiente función en R permite evaluar las probabilidades anteriores:

```
probs <- function(n,pDet_C,pDet_L,pC){
  result <- NULL
  for (m in 1:n){
    pDL_C=sum(dbinom(0:(m-1),n,pDet_C))
    pDC_C=1-pDL_C
    pDC_L=sum(dbinom(m:n,n,pDet_L))
    pDL_L=1-pDC_L
    # Prob. total
    pDL <- pDL_C*pC+pDL_L*(1-pC)
    pDC <- pDC_C*pC+pDC_L*(1-pC)
    # Bayes
    pC_DL <- pDL_C*pC/pDL
    pL_DC <- pDC_L*(1-pC)/pDC
    probs <- data.frame(m,pDL_C,pDC_L,pC_DL,pL_DC)
    result <- rbind(result,probs)
  }
  result
}
```

Mejora del protocolo de detección de contaminantes

Regla de decisión: Si $N_D \geq m$ se decide D^+ (hay contaminación), y si $N_D < m$ se decide D^- (la playa está limpia).

Si la muestra es de tamaño $n = 5$ la función calcula las probabilidades para cada valor de m:

```
probs(n=5,pDet_C=0.99,pDet_L=0.10,pC=0.03) %>% flextable() %>%
  fontsize(size=15)
```

m	pDL_C	pDC_L	pC_DL	pL_DC
1	0.0000000001	0.40951	0.000000000005237656	0.929779341
2	0.0000000496	0.08146	0.000000001670064032	0.724811552
3	0.0000098506	0.00856	0.000000307288023825	0.216777294
4	0.0009801496	0.00046	0.000030326936217431	0.014669527
5	0.0490099501	0.00001	0.001513492678206095	0.000339881

Otro ejemplo: Si $P(D|C) = 0.75$ y $P\left(D\middle|\overline{C}\right) = 0.15$, con $P(C) = 0.03$

```
probs(n=10,pDet_C=0.75,pDet_L=0.15,pC=0.03) %>% flextable() %>%
  fontsize(size=11)
```

m	pDL_C	pDC_L	pC_DL	pL_DC
1	0.0000009536743	0.803125595659277	0.0000001498167	0.962918674618
2	0.0000295639038	0.455700176234473	0.0000016798572	0.936446365403
3	0.0004158020020	0.179803519632422	0.0000156787482	0.853287794121
4	0.0035057067871	0.049969798878516	0.0001141137918	0.618520701874
5	0.0197277069092	0.009874090998633	0.0006158403833	0.245674323184
6	0.0781269073486	0.001383235212305	0.0024138024986	0.046270134201
7	0.2241249084473	0.000134579949609	0.0068849007151	0.005577121920
8	0.4744071960449	0.000008665133203	0.0144603450217	0.000532776265
9	0.7559747695923	0.000000332535059	0.0228465041604	0.000044058941
10	0.9436864852905	0.000000005766504	0.0283585037278	0.000003310922

Mejora del protocolo de detección de contaminantes

En resumen: Si la realización de un único análisis (toma de una única muestra) lleva asociados riesgos de error importantes (**siempre hay dos tipos de error**), deberemos fijar una regla de decisión (tamaño de muestra n más punto de corte m) que produzca riesgos de error aceptables.

El riesgo 0 no es posible en la práctica.

Ejemplo 4: Pesca en Canarias

La siguiente tabla muestra qué porcentaje de todos los barcos pesqueros que operan en Canarias se encuentra matriculado en cada una de las islas:

EH	FV	GC	LG	LP	LZ	TF
7%	15%	19%	3%	8%	12%	36%

A su vez, la siguiente tabla recoge, para cada isla, qué proporción de los barcos pesqueros matriculados en ella utiliza la nasa como arte de pesca:

EH	FV	GC	LG	LP	LZ	TF
24%	47%	74%	22%	77%	68%	69%

¿Cuál es la proporción total de barcos pesqueros que emplean nasas en Canarias? Y de los que usan nasas, ¿qué proporción está matriculada en Gran Canaria?

Ejemplo 4: Pesca en Canarias

La resolución de la primera pregunta es una simple aplicación del teorema de la probabilidad total. La probabilidad de que un barco elegido al azar esté matriculado en cada isla es:

$$P(EH) = 0.07 \quad P(FV) = 0.15 \quad P(GC) = 0.19 \quad P(LG) = 0.03$$

$$P(LP) = 0.08 \quad P(LZ) = 0.12 \quad P(TF) = 0.36$$

A su vez, la probabilidad de que utilize nasas depende de la isla en que esté matriculado:

$$P(N|EH) = 0.24 \quad P(N|FV) = 0.47 \quad P(N|GC) = 0.74$$

$$P(N|LG) = 0.22 \quad P(N|LP) = 0.77 \quad P(N|LZ) = 0.68$$

$$P(N|TF) = 0.69$$

Ejemplo 4: Pesca en Canarias

Por tanto la probabilidad total de usar nasas es:

$$P(N) = \sum_{k=1}^{7} P(N|Isla_k)\,P(Isla_k) = P(N|EH)\,P(EH) + P(N|FV)\,P(FV$$

$$+P(N|GC)\,P(GC) + P(N|LG)\,P(LG) + P(N|LP)\,P(LP) +$$

$$+P(N|LZ)\,P(LZ) + P(N|TF)\,P(TF) = 0.6261$$

es decir, usan nasas el 62.61% de los barcos pesqueros matriculados en Canarias

Ejemplo 4: Pesca en Canarias

Ahora, para calcular qué proporción de los que usan nasas se encuentran matriculados en Gran Canaria, aplicamos el teorema de Bayes:

$$P\left(GC|\,N\right) = \frac{P\left(N|\,GC\right)P\left(GC\right)}{P\left(N\right)} = \frac{0.74 \cdot 0.19}{0.6261} = 0.2246$$

Por tanto, de todos los barcos que usan nasas, el 22.46% se encuentra matriculado en Gran Canaria.