Trabajo práctico 5: Probabilidad

AL TERMINAR ESTA GUÍA

- Calcular probabilidades con las reglas de la suma y del producto, y probabilidades condicionales a partir de una tabla.
- Distinguir eventos independientes de eventos mutuamente excluyentes.
- Resolver problemas de probabilidad total y de Bayes con una tabla de frecuencias naturales y con la fórmula.
- Contar con permutaciones, variaciones y combinaciones, y estimar una probabilidad por simulación.

PARA EL DOCENTE

Los ejercicios 1 a 3 son para el encuentro práctico; el 2 es el más importante de la guía y conviene resolverlo primero con la tabla de 10.000 personas y recién después con la fórmula. El 4 necesita computadora: sirve para un encuentro virtual o como tarea.

El 5 y el 6 son para discutir en grupos. Antes del 4 se puede preguntar en el curso si hay dos personas que cumplen el mismo día: con 40 estudiantes, casi siempre hay.

EJERCICIO 1 Trabajar y aprobar

A mano

En una carrera se registró, para los 280 estudiantes que rindieron el primer parcial de Estadística, si trabajan y si aprobaron:

	Aprobó	No aprobó
Trabaja	32	44
No trabaja	131	73

Se elige un estudiante al azar. Sean T el evento «trabaja» y A el evento «aprobó».

a) ¿Cuál es el experimento aleatorio y cuál el espacio muestral? Calculá P(T) y P(A).

RESOLUCIÓN

El experimento es elegir un estudiante al azar; el espacio muestral, los 280 estudiantes, todos con la misma probabilidad de salir. Con resultados igualmente probables, la probabilidad es casos favorables sobre casos posibles (la definición clásica):

$$P(T) = \frac{76}{280} = 0{,}271 \qquad P(A) = \frac{163}{280} = 0{,}582$$

b) Calculá P(T y A) y P(T o A). Usá la regla de la suma y comprobá el resultado contando en la tabla.

> RESOLUCIÓN
>
> $$P(T \cap A) = \frac{32}{280} = 0{,}114$$
>
> $$P(T \cup A) = P(T) + P(A) - P(T \cap A) = 0{,}271 + 0{,}582 - 0{,}114 = 0{,}739$$
>
> Contando: trabajan o aprobaron todos menos los que ni trabajan ni aprobaron, 280 − 73 = 207, y 207/280 = 0,739. Se resta la intersección porque los 32 que trabajan y aprobaron se contaron dos veces, una en T y otra en A.

c) Calculá la probabilidad de aprobar entre los que trabajan y entre los que no trabajan. ¿Son independientes los eventos T y A?

> RESOLUCIÓN
>
> $$P(A \mid T) = \frac{P(A \cap T)}{P(T)} = \frac{32}{76} = 0{,}421 \qquad P(A \mid \bar{T}) = \frac{131}{204} = 0{,}642$$
>
> No son independientes: saber que el estudiante trabaja cambia la probabilidad de que haya aprobado, de 0,582 a 0,421. Si fueran independientes, P(A | T) sería igual a P(A), y lo mismo P(T ∩ A) = P(T) · P(A) = 0,158, que no es 0,114.

d) ¿Son T y A mutuamente excluyentes? ¿Qué diferencia hay entre «excluyentes» e «independientes»?

> RESOLUCIÓN
>
> No: hay 32 estudiantes que trabajan y aprobaron, así que los dos eventos pueden ocurrir juntos. Excluyentes significa que no pueden ocurrir a la vez (P(T ∩ A) = 0); independientes, que uno no cambia la probabilidad del otro. Son ideas distintas, y casi opuestas: dos eventos excluyentes con probabilidad positiva nunca son independientes, porque si ocurre uno, el otro pasa a tener probabilidad cero.

e) Se eligen dos estudiantes al azar, sin reposición. ¿Cuál es la probabilidad de que los dos trabajen?

> RESOLUCIÓN
>
> $$P(T_1 \cap T_2) = P(T_1) \cdot P(T_2 \mid T_1) = \frac{76}{280} \cdot \frac{75}{279} = 0{,}0730$$
>
> Es la regla del producto con probabilidad condicional: una vez elegido el primero, queda un estudiante que trabaja menos y un estudiante menos en total. Con reposición sería P(T)², apenas distinto con tantos estudiantes.

EJERCICIO 2 Una prueba que da positivo

A mano

Una enfermedad afecta al 1 % de la población adulta. Para detectarla se usa una prueba rápida que da positivo en el 96 % de las personas enfermas (sensibilidad) y da negativo en el 95 % de las sanas (especificidad).

Una campaña de salud le hace la prueba a toda la población adulta de una ciudad.

a) Imaginá 10.000 personas y armá la tabla: enfermas y sanas, por resultado positivo y negativo.

RESOLUCIÓN

	Positivo	Negativo	Total
Enfermas	96	4	100
Sanas	495	9.405	9.900
Total	**591**	**9.409**	**10.000**

Se arma de afuera hacia adentro: el 1 % de 10.000 son 100 enfermas; el 96 % de ellas da positivo (96); el 95 % de las 9.900 sanas da negativo (9.405), y el resto de cada fila sale por diferencia. Pensar en cantidades de personas en lugar de probabilidades —frecuencias naturales— hace el problema mucho más fácil de razonar.

b) ¿Cuál es la probabilidad de que una persona elegida al azar dé positivo? Calculala con la tabla y con la fórmula de probabilidad total.

RESOLUCIÓN

Con la tabla: 591 positivos de 10.000. Con la fórmula, recorriendo las dos ramas del árbol —enferma y positiva, sana y positiva—:

$$P(+) = P(+ \mid E)\,P(E) + P(+ \mid \bar{E})\,P(\bar{E}) = 0{,}96 \cdot 0{,}01 + 0{,}05 \cdot 0{,}99 = 0{,}0591$$

c) Una persona da positivo. ¿Cuál es la probabilidad de que esté enferma? ¿Te sorprende?

RESOLUCIÓN

$$P(E \mid +) = \frac{P(+ \mid E)\,P(E)}{P(+)} = \frac{0{,}96 \cdot 0{,}01}{0{,}0591} = \frac{96}{591} = 0{,}162$$

Solo 16,2 %: de cada 100 positivos, unos 16 están enfermos. Sorprende porque la prueba parece muy buena, pero la enfermedad es rara: los 495 falsos positivos que salen de las 9.900 personas sanas superan a los 96 verdaderos positivos. Es el valor predictivo positivo, y es lo que le importa a quien recibe el resultado.

d) Si una persona da negativo, ¿cuál es la probabilidad de que esté sana?

RESOLUCIÓN

$$P(\bar{E} \mid -) = \frac{9.405}{9.409} = 0{,}9996$$

Un negativo es muy confiable: de 9.409 negativos, solo 4 son personas enfermas que la prueba no detectó. Por eso una prueba así sirve para descartar la enfermedad y, si da positivo, se confirma con un estudio más preciso.

e) La misma prueba se usa en un consultorio de especialistas, donde el 20 % de los pacientes que llegan tiene la enfermedad. ¿Cuánto vale ahí la probabilidad de estar enfermo si da positivo? ¿Por qué cambia, si la prueba es la misma?

RESOLUCIÓN

$$P(E \mid +) = \frac{0{,}96 \cdot 0{,}20}{0{,}96 \cdot 0{,}20 + 0{,}05 \cdot 0{,}80} = 0{,}828$$

Pasa de 16,2 % a 82,8 %. La sensibilidad y la especificidad son propiedades de la prueba; los valores predictivos dependen además de la prevalencia, de cuán común es la enfermedad entre las personas a las que se les hace. Por eso el mismo positivo significa cosas distintas en un tamizaje masivo y en un consultorio al que llegan pacientes con síntomas.

EJERCICIO 3 La comisión y la clave

A mano

Un curso tiene 12 estudiantes: 7 mujeres y 5 varones. Hay que formar una comisión de tres para representar al curso.

a) ¿De cuántas maneras distintas se puede formar la comisión, si los tres integrantes tienen el mismo rol?

RESOLUCIÓN

$$\binom{12}{3} = \frac{12!}{3!\,9!} = \frac{12 \cdot 11 \cdot 10}{3 \cdot 2 \cdot 1} = 220$$

No importa el orden —la comisión Ana, Beto y Caro es la misma que Caro, Ana y Beto— y nadie se repite: son combinaciones.

b) ¿Y si los tres cargos son distintos: delegado, suplente y secretario?

RESOLUCIÓN

$$V_{12,3} = 12 \cdot 11 \cdot 10 = 1.320$$

Ahora el orden importa: hay 12 candidatos para delegado, 11 para suplente y 10 para secretario (principio de multiplicación). Cada comisión de la consigna anterior se puede repartir de 3! = 6 maneras: 220 · 6 = 1.320.

c) ¿Cuántas comisiones de tres (sin cargos) tienen al menos una mujer?

RESOLUCIÓN

$$\binom{12}{3} - \binom{5}{3} = 220 - 10 = 210$$

«Al menos una» incluye una, dos o tres mujeres: tres cuentas. Es más corto contar lo contrario —ninguna mujer, o sea tres varones— y restarlo del total.

d) Si la comisión se sortea, ¿cuál es la probabilidad de que salgan tres mujeres? ¿Y de que haya al menos un varón?

RESOLUCIÓN

$$P(\text{tres mujeres}) = \frac{\binom{7}{3}}{\binom{12}{3}} = \frac{35}{220} = 0{,}159 \qquad P(\text{al menos un varón}) = 1 - 0{,}159 = 0{,}841$$

El sorteo hace que todas las comisiones sean igualmente probables, y entonces la probabilidad es casos favorables sobre casos posibles, contados con combinaciones.

e) La clave del aula virtual es de cuatro dígitos. ¿Cuántas claves hay? ¿Cuántas no repiten ningún dígito? Si alguien elige una clave al azar, ¿qué probabilidad hay de que no tenga dígitos repetidos?

RESOLUCIÓN

$10^4 = 10.000 \qquad V_{10,4} = 10 \cdot 9 \cdot 8 \cdot 7 = 5.040 \qquad P = \frac{5.040}{10.000} = 0{,}504$

Con repetición, cada lugar tiene 10 opciones; sin repetición, una menos en cada lugar. Casi la mitad de las claves al azar tienen algún dígito repetido: más de lo que la intuición suele decir.

EJERCICIO 4 El problema del cumpleaños

Con software

En un curso de 40 estudiantes, ¿es probable que dos cumplan años el mismo día? Antes de calcular nada, anotá tu estimación. Suponé que los 365 días son igualmente probables y olvidate del 29 de febrero.

Se puede usar R, Python o Excel. En R, sample(365, 23, replace = TRUE) sortea 23 cumpleaños y any(duplicated(x)) dice si hay alguno repetido.

a) Calculá la probabilidad exacta de que en un grupo de 23 personas al menos dos cumplan el mismo día. Pista: es más fácil calcular la probabilidad de que todos cumplan en días distintos.

RESOLUCIÓN

$$P(\text{todos distintos}) = \frac{365}{365} \cdot \frac{364}{365} \cdots \frac{343}{365} = \frac{V_{365,23}}{365^{23}} = 0{,}4927$$

$$P(\text{al menos dos iguales}) = 1 - 0{,}4927 = 0{,}5073$$

Con 23 personas ya es más probable que haya un cumpleaños compartido que no. Sorprende porque uno piensa en «alguien que cumpla el mismo día que yo», pero cuentan todos los pares: con 23 personas hay 253 pares distintos. En R: 1 - prod((365:343) / 365).

b) Simulá 10.000 grupos de 23 personas y contá en qué proporción hay al menos un cumpleaños repetido. Compará con el resultado exacto.

RESOLUCIÓN

En R: mean(replicate(10000, any(duplicated(sample(365, 23, replace = TRUE))))). En Python: np.mean([len(set(rng.integers(0, 365, 23))) < 23 for _ in range(10000)]).

El resultado cambia en cada corrida, pero tiene que andar cerca de 0,507: con 10.000 repeticiones, casi siempre entre 0,497 y 0,517. La simulación estima la probabilidad con su definición frecuencial: la proporción de veces que ocurre el evento en muchas repeticiones del experimento.

c) Repetí la simulación con 10, 100, 1.000 y 10.000 grupos, y anotá la proporción que da cada vez. ¿Qué pasa a medida que crece la cantidad de repeticiones?

RESOLUCIÓN

Con 10 grupos la proporción puede dar cualquier cosa entre 0,2 y 0,8; con 100, se mueve unas décimas; con 10.000, se queda a una centésima de la probabilidad exacta. Es la ley de los grandes números: la frecuencia relativa se acerca a la probabilidad al aumentar las repeticiones. Lo que no dice la ley es que los resultados «se compensen»: cada grupo es independiente de los anteriores.

d) Calculá la probabilidad para un curso de 10 y de 40 estudiantes. ¿Desde cuántas personas la probabilidad supera el 50 %? ¿Y el 90 %?

RESOLUCIÓN

Con 10 personas, 0,117; con 40, 0,891: en un curso de 40 es casi seguro que haya un cumpleaños compartido. La probabilidad pasa el 50 % con 23 personas y el 90 % con 41.

En R alcanza con una línea para todos los tamaños: n ← 1:60; p ← sapply(n, function(k) 1 - prod((365:(365 - k + 1)) / 365)). Graficarla muestra que crece rápido entre 10 y 50 personas.

EJERCICIO 5 Razonamientos que fallan

Para discutir

Cada frase contiene un error de razonamiento sobre probabilidades. Encontralo y explicá cómo se razona bien.

a) «En la ruleta salió rojo siete veces seguidas. Ahora tiene que salir negro.»

RESOLUCIÓN

La falacia del jugador. La ruleta no tiene memoria: cada tirada es independiente de las anteriores, y la probabilidad de negro sigue siendo la misma que antes. La ley de los grandes números no funciona compensando rachas, sino diluyéndolas en muchísimas tiradas.

b) «O me toca la lotería o no me toca: tengo un 50 % de chances.»

RESOLUCIÓN

Que haya dos resultados posibles no los hace igualmente probables. La definición clásica —favorables sobre posibles— solo vale si todos los resultados tienen la misma probabilidad. Con una lotería de 14 millones de combinaciones, la probabilidad de ganar con una es una en 14 millones.

c) «Jugué 1-2-3-4-5-6 en la lotería. Mi amigo dice que es una tontería, porque esa combinación nunca va a salir.»

RESOLUCIÓN

Cualquier combinación concreta tiene exactamente la misma probabilidad que 1-2-3-4-5-6. Parece menos probable porque «parece ordenada», y las combinaciones «desordenadas» son tantas que se las piensa como un solo resultado. Lo único que cambia es que, si sale, se comparte el premio con más gente: hay quienes la juegan.

d) «El 90 % de las personas con esta enfermedad tiene el síntoma. Tengo el síntoma, así que tengo un 90 % de probabilidad de estar enfermo.»

RESOLUCIÓN

Confunde P(síntoma | enfermo) con P(enfermo | síntoma), que pueden ser muy distintas. Si la enfermedad es rara y el síntoma es común —un dolor de cabeza, por ejemplo—, la mayoría de los que tienen el síntoma no están enfermos. Es el mismo razonamiento del ejercicio de la prueba de tamizaje: hace falta Bayes y la prevalencia.

EJERCICIO 6 ¿De dónde sale cada probabilidad?

Para discutir

Para cada probabilidad, decidí si sale de la definición clásica (casos favorables sobre posibles), de la frecuencial (lo que pasó en muchas repeticiones) o de la subjetiva (el grado de confianza de alguien).

a) La probabilidad de sacar un as de un mazo de 40 cartas españolas es 4/40.

RESOLUCIÓN

Clásica: 40 resultados igualmente probables, 4 favorables. No hace falta sacar ninguna carta para saberla.

b) La probabilidad de que un recién nacido sea varón es aproximadamente 0,51.

RESOLUCIÓN

Frecuencial: sale de las estadísticas vitales, de millones de nacimientos registrados. No es 0,5 —los dos sexos no son igualmente probables—, así que la definición clásica daría un número equivocado.

c) El pronóstico dice que mañana hay 70 % de probabilidad de lluvia.

RESOLUCIÓN

Es frecuencial en su origen: entre los días con condiciones atmosféricas parecidas, en el 70 % llovió, y los modelos del pronóstico se calibran así. Mañana es un día único que no se repite, y por eso tiene también algo de subjetiva.

d) Un comentarista dice que su equipo tiene 30 % de probabilidad de salir campeón este año.

RESOLUCIÓN

Subjetiva: el campeonato de este año ocurre una sola vez y no hay resultados igualmente probables. Es un grado de confianza, que puede estar mejor o peor fundado, y otra persona con la misma información podría dar otro número.

---

estadistica.ar/docentes/trabajos-practicos/semana-5 · Las lecturas de cada tema y la hoja de fórmulas están en el sitio.