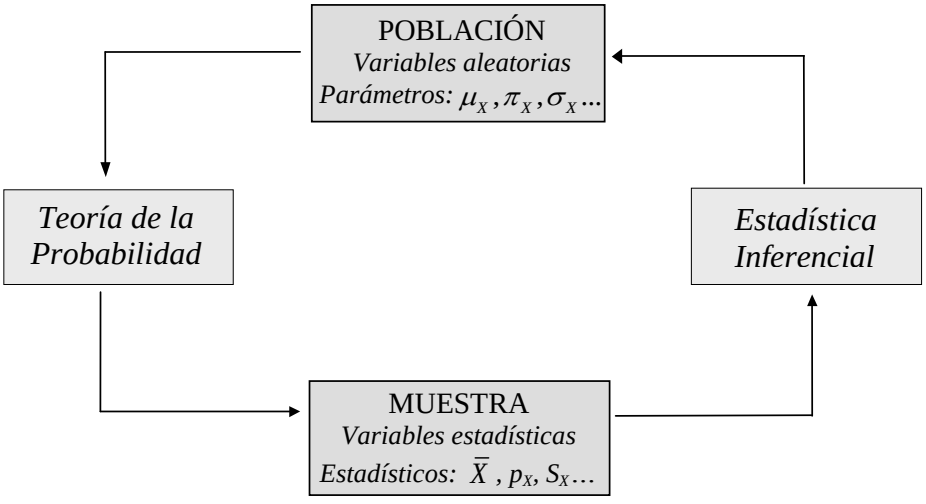
7 – Uso de la probabilidad en la investigación psicológica

1. Teoría de la Probabilidad

2. Variables aleatorias

• La importancia de la Teoría de la Probabilidad en el ámbito de la estadística se deriva del hecho de constituir ésta uno de los pilares teóricos sobre los que se asienta el desarrollo y aplicación de la Estadística Inferencial. Así, tal como se ilustra en la figura que se muestra más abajo, cuando de una o más variables conocemos en la población sus características (tendencia central, dispersión, asociación...), es la Teoría de la Probabilidad la que nos va a proporcionar las herramientas que nos permitan establecer predicciones de las características que esas variables adoptarán en una muestra de sujetos extraída al azar de esa población. En sentido inverso, la estadística inferencial -basándose en el conocimiento desarrollado por la Teoría de la Probabilidad en ese camino de la población a la muestra- ha establecido las bases para trazar el camino opuesto: inferir, a partir de los datos en una o más variables de una muestra, cómo serán las características (tendencia central, dispersión, asociación...) de esas variables en la población a la que esa muestra representa.



1. Teoría de la Probabilidad

• Ante un evento de resultado incierto, el campo de conocimientos de la Teoría de la Probabilidad ha dirigido sus esfuerzos a determinar el grado en qué puede ocurrir cualquiera de los resultados posibles [sucesos] que se pueden derivar de la realización de tal evento incierto [experimento aleatorio] -entre corchetes, las expresiones más utilizadas a nivel académico.

> **Ejemplos** de evento incierto: (1) el lanzamiento de una moneda (sucesos posibles: que salga cara; que salga cruz); mi estado de salud durante el próximo mes (sucesos posibles: bueno; malo; regular); la práctica religiosa de un estudiante de la Universidad elegido al azar (sucesos posibles: ninguna; católica; protestante; etc.); el *CI* de ese mismo estudiante (sucesos posibles: que sea igual a 120; que sea igual a 85; que sea...). En este último caso, los sucesos se podrían expresar, no en forma de *sucesos elementales*, sino de *sucesos compuestos*, por ejemplo: que sea menor de 110; que sea mayor o igual a 110; que esté entre 89 y 120, etc.

• El esfuerzo de la Teoría de la Probabilidad por determinar el grado en que puede ocurrir uno cualquiera de los sucesos asociado a un determinado experimento aleatorio se ha concretado en la asignación de un valor numérico que refleje el grado en que es previsible la ocurrencia de ese suceso. A este valor numérico se le conoce como probabilidad (*P*) y puede, por convención, oscilar entre 0 y 1 (0: probabilidad nula; 1: probabilidad segura). Así, para un suceso *i* de un experimento aleatorio *X*, la anterior propiedad se expresa como:

$$0 \leq P(X_i) \leq 1$$

Otra propiedad importante de las probabilidades es que, para los distintos (*n*) sucesos elementales asociados a un experimento aleatorio, la suma de sus probabilidades será igual a 1:

$$\sum_{i=1}^{n} P(X_i) = 1$$

• A continuación se van a describir 3 enfoques en la estimación de las probabilidades asociadas a los resultados posibles de un evento incierto o experimento aleatorio. Dado que normalmente estos enfoques lo que permiten obtener son estimaciones, no los verdaderos valores de probabilidad, haremos referencia a estos valores estimados con el símbolo *P'*, mientras que para el verdadero valor de probabilidad se reserva el símbolo *P*.

(1) Enfoque subjetivo: supone estimar la probabilidad de un suceso en función del grado de confianza personal que se tiene acerca de la ocurrencia del mismo, ya venga esa confianza determinada por nuestra experiencia vital, por nuestras convicciones personales o creencias, o por cualquier otra

fuente sobre la que se base el conocimiento que tenemos de nuestro entorno. Se trata del procedimiento más utilizado desde siempre en la práctica a la hora de estimar probabilidades, especialmente, cuando no se tienen ciertas nociones sobre otras aproximaciones al cálculo de probabilidades, o bien, cuando aplicar éstas resulte poco operativo. Por ejemplo, cuando me asomo a la ventana antes de salir de casa y veo el cielo, realizo una estimación de la probabilidad de que llueva durante el día. Y como consecuencia de esa estimación, y dependiendo de lo que me importe mojarme, decido qué ponerme o si coger un paraguas o no. En realidad, hacemos este tipo de estimaciones subjetivas de probabilidad en múltiples situaciones, aunque no siempre de forma del todo consciente, constituyendo un elemento determinante de las decisiones que finalmente tomamos.

> **Ejemplos** de estimación subjetiva de probabilidad: (1) a la hora de estimar la probabilidad de que al lanzar dos dados salgan en ambos un seis, muchas personas realizarían una estimación subjetiva de la misma pues, aunque existen otras aproximaciones más precisas a la hora de realizar esa estimación, su aplicación es desconocida para muchos; (2) también las personas suelen realizar estimaciones subjetivas de la probabilidad de que les toque el 'gordo' en un sorteo de lotería -en general, muy al alza- y, curiosamente, suelen ser estimaciones diferentes en función del número considerado; (3) también es habitual realizar estimaciones subjetivas de la probabilidad respecto al resultado de un partido, por ejemplo, de que gane el Valencia CF en su partido del próximo fin de semana.

(2) Enfoque clásico o a priori: consiste en estimar la probabilidad de un suceso (X_i) como la razón entre los resultados favorables a ese suceso y el número total de resultados posibles que se pueden dar en la realización del experimento aleatorio.

$$P'(X_i) = \frac{n^o\, de\, resultados\ favorables}{n^o\, de\, resultados\ posibles}$$

> **Ejemplo**: ¿cuál es la probabilidad de que al lanzar un dado salga un 5?
>
> $$P'(X_i = 5) = \frac{1}{6} = 0{,}167$$

Ejercicio 1: ¿cuál es la probabilidad de que al lanzar un dado salga un 3?; ¿y de que salga número par?; ¿y de que al lanzar dos dados, la suma de los puntos dé igual a 7?; ¿y de que en la lotería de Navidad toque el *gordo* en el número al que juego?

Hay que subrayar que, a la hora de calcular una probabilidad, la aplicación del enfoque clásico asume el conocido como principio de indiferencia, esto es, que la probabilidad de ocurrencia de todos los sucesos es la misma. Si se cumple este supuesto en la realización de un determinado experimento

aleatorio, entonces podremos decir que las estimaciones realizadas de acuerdo a esta aproximación serán los verdaderos valores de probabilidad. Sin embargo, el cumplimiento de este principio que asume que los sucesos son equiprobables resulta difícil de aceptar en muchas situaciones en la práctica. Por ejemplo, si aplicamos la aproximación clásica a la hora de estimar la probabilidad de que un estudiante elegido al azar de la Universidad su estado civil sea viudo/a, nos daría igual a ¼, un resultado poco creíble pero que ha venido motivado por realizar la estimación no cumpliéndose en este caso el principio de indiferencia. Por experiencia, bien sabemos que los sucesos soltero/a, casado/a, separado/a y viudo/a no son en absoluto equiprobables.

En algunos casos sí que se puede asumir el cumplimiento de este principio –por ejemplo, en juegos de azar–, pero en otros muchos casos se puede tener serias dudas acerca de la satisfacción del mismo, lo cual cuestionaría la aplicación de este enfoque.

(3) Enfoque frecuencialista, a posteriori o estadístico: dado un suceso X_i asociado a la realización de un determinado experimento aleatorio, la estimación de la probabilidad de X_i a partir de este enfoque se basa en la repetición de una gran cantidad de veces del experimento aleatorio en las mismas condiciones, para así obtener la razón entre el nº de veces que ha ocurrido ese suceso (n_i) y el nº de repeticiones del experimento (n):

$$P'(X_i) = \frac{n_i}{n}$$

Ejercicio 2: ¿Cómo se estimaría la probabilidad, de acuerdo a esta aproximación, de que salga un 3 en el lanzamiento de un dado? ¿Y del resto de sucesos planteados en el ejercicio 1?

De acuerdo al enfoque frecuencialista, cuanto mayor sea el número de repeticiones del experimento aleatorio, más cercano será el valor de probabilidad estimado ($P'(X_i)$) al verdadero valor de probabilidad $P(X_i)$. En términos matemáticos:

$$P(X_i) = \lim_{n\to\infty} \frac{n_i}{n}$$

> **Ejemplo**: Si lanzamos una moneda 10 veces a fin de estimar la probabilidad de que salga cara, la probabilidad estimada podría ser, por ejemplo, P'(cara) = 0,6 si nos salieran 6 caras y 4 cruces en esos 10 lanzamientos. Sin embargo, a medida que aumenta el número de lanzamientos (idealmente, hasta infinito) esta estimación se irá acercando a la probabilidad verdadera. Se supone que ese valor será igual a 0,5, pero no tiene por qué necesariamente ser así ya que la moneda podría tener algún tipo de curvatura o ser más pesada por alguno de los dos lados.

Si nos fijamos en la fórmula de la aproximación frecuencialista, la estimación de la probabilidad de un suceso se corresponde con la fórmula de la frecuencia relativa o proporción (p_i) que vimos al construir una distribución de frecuencias:

$$P'(X_i) = p_i$$

En realidad, cualquiera de las variables vistas en los ejemplos de los temas precedentes puede contemplarse como la repetición de un experimento aleatorio concreto, pues consiste en la medición de un atributo determinado en múltiples ocasiones -tantas como diferentes sujetos sean medidos-, sin tener certidumbre a priori de cuál van a ser los valores resultantes (sucesos) de esas mediciones. Y de acuerdo a la aproximación frecuencialista de la probabilidad, las frecuencias relativas que se obtengan al calcular la distribución de frecuencias de esa variable, representarán la estimación de las probabilidades asociadas a esos valores de la variable (experimento aleatorio).

> **Ejemplo**: Queremos estimar la probabilidad de estar casado (X_i = casado) dentro de los estudiantes de la UVEG y disponemos de una muestra de 500 estudiantes de dicha universidad:
> - Experimento aleatorio: obtener información del estado civil de un estudiante de la UVEG.
> - Repeticiones del experimento aleatorio: se pregunta a 500 estudiantes (n = 500).
> - Nº de ocurrencias del suceso de interés: nº de estudiantes casados en esa muestra, supongamos que son 60 (n_{casado} = 60).
> - $P'(X$ = casado) = 60/500 = 0,12. Esta estimación se aproximará a la probabilidad verdadera cuanto mayor sea el nº de repeticiones, en este caso, el tamaño de la muestra. Se considerarán como los verdaderos valores de probabilidad en el caso en que se cuente con datos para todos los elementos de la población.

Ejercicio 3 (adaptado a partir de ejemplo de Barón-López, 2005): Se ha repetido en 1000 ocasiones el experimento de elegir a una mujer de la población española de mujeres de entre 45 y 55 años, obteniéndose datos de las variables "Nivel de masa ósea" [*NO*: Normal; *ON*: Osteopenia; *OR*: Osteoporosis (según clasificación de la *OMS*)] y "Haber pasado la menopausia" [*N*: No; *S*: Sí]. Los datos obtenidos se muestran resumidos en la siguiente tabla de contingencia:

		MENOPAUSIA		
		NO	SI	Total
CLASIFICACION OMS	NORMAL	189	280	469
	OSTEOPENIA	108	359	467
	OSTEOPOROSIS	6	58	64
Total		303	697	1000

A partir de los datos recogidos, contéstese a las siguientes cuestiones (entre corchetes aparece el equivalente de la cuestión, expresada de forma simbólica):

a) ¿Cuál es la probabilidad (estimada) de que una mujer (extraída al azar de la población española de mujeres de entre 45 y 55 años) tenga osteoporosis? [$P'(OR)$]

b) ¿Y de que no haya pasado la menopausia? [$P'(N)$]

c) ¿Y de que tenga osteopenia u osteoporosis? [$P'(ON \cup OR)$]

d) ¿Y de que no haya pasado la menopausia y tenga osteoporosis? [$P'(N \cap OR)$]

e) ¿Y de que tenga osteoporosis si sabemos que no ha pasado la menopausia? [$P(OR|N)$]

f) Conociendo que una mujer tiene osteopenia ¿cuál es la probabilidad estimada de que haya pasado la menopausia? [$P'(S|ON)$]

En este ejercicio se plantea la aplicación práctica de alguno de los teoremas básicos de la probabilidad (más detalles sobre los mismos en, por ejemplo, Botella y cols. (2001, tema 12):

- probabilidad de la intersección de dos sucesos: $P(A \cap B)$
- probabilidad de la unión de dos sucesos: $P(A \cup B)$
- probabilidad condicional: $P(B|A)$ o $P(A|B)$

2. Variables aleatorias

• Frente al concepto de variable estadística, el concepto de variable aleatoria supone contar con información de la probabilidad asociada a cada una de las modalidades de la variable, lo cual implica contar con datos para la población pues, en otro caso, lo que tendríamos serían frecuencias relativas, esto es, estimaciones de las probabilidades, no las probabilidades en sí.

> **Ejemplo:** Si medimos la variable "Estado civil" en toda la población de estudiantes de la *UVEG* (N = 45000) y obtenemos que 350 son viudos/as, entonces la frecuencia relativa correspondiente a la modalidad 'ser viudo/a' (p_{viudo}= 350/45000 = 0,008) será precisamente la probabilidad de 'ser viudo' (P_{viudo}) en la población de estudiantes de la *UVEG*, y no una estimación de la misma. Análogamente, si obtenemos las probabilidades asociadas a las otras modalidades (sucesos) de la variable "Estado civil", tendremos la distribución de probabilidad asociada a esta variable aleatoria en la citada población. Sea, por ejemplo, la siguiente:

X_i	$P(X_i)$
soltero/a	0,884
casado/a	0,105
separado/a	0,009
viudo/a	0,002
	1,00

• La distribución de probabilidad de una variable aleatoria –de forma análoga a la distribución de frecuencias de una variable estadística– consiste en la correspondencia entre los distintos valores que toma la variable y las probabilidades asociadas a esos valores.

– A esa correspondencia entre las modalidades de una variable y sus probabilidades (i.e., distribución de probabilidad) se le suele llamar función de probabilidad en el caso de tratarse de una variable aleatoria discreta (variables categóricas, ordinales o cuantitativas discretas) y función de densidad de probabilidad en el caso de que sea continua (variables cuantitativas continuas).

Ejemplo: función de probabilidad correspondiente a la variable "Nº de contratos laborales en los 2 últimos años" para la población de personas en edad laboral de la comarca del Camp de Morvedre:

X_i	$P(X_i)$
0	0,08
1	0,31
2	0,35
3	0,18
4	0,07
5	0,01
	1

– Y se llama función de distribución a la que hace corresponder a cada valor de la variable, la probabilidad de que se dé un valor como ese o inferior (P_a), concepto análogo al de frecuencia relativa acumulada. Este concepto no es aplicable si la variable es categórica, pues carece de sentido el concepto de probabilidad acumulada para este tipo de variable.

Ejemplo: función de distribución para a la variable "Nº de contratos laborales…":

X_i	$P_a(X_i)$
0	0,08
1	0,39
2	0,74
3	0,92
4	0,99
5	1

• La distribución de probabilidad de una variable no suele ser conocida dado que, con frecuencia, no resulta viable recoger datos de todos los elementos de la población de interés para una determinada variable. Una aproximación a la misma consiste en estimar las probabilidades correspondientes a partir de los datos recogidos para una muestra de esa población, aplicando para ello la aproximación frecuencialista al cálculo de las probabilidades. Otra vía de aproximación a la distribución de probabilidad de una variable consiste en asumir, atendiendo a razones sustantivas o a la experiencia práctica acumulada, que dicha variable se distribuye de acuerdo a algún modelo teórico de

características conocidas, tal como algunos de los que se presentarán en la siguiente sección (distribución normal, distribución binomial, distribución *t* de Student…).

• Cuando se obtiene un índice cualquiera -por ejemplo, la media- a partir de la distribución de probabilidad de una variable, al valor resultante se le denomina parámetro –si fuera a partir de una distribución de frecuencias, sería un estadístico.

• Se suelen utilizar letras griegas minúsculas para representar a los parámetros. Así, por ejemplo, dada una variable *X*, se utiliza μ_X para representar la media, σ_X para la desviación típica, π_X para la proporción... Es lo mismo para el caso de los índices estadísticos bivariados, por ejemplo, σ_{XY} para la covarianza, ρ_{XY} para el coeficiente de correlación de Pearson, β_0 para la constante de la ecuación de regresión, β_1 para la pendiente... Hay algún caso especial, siendo el más notable el de la media aritmética que, como parámetro, aparece también representada como *E(X)* y denominada, de forma alternativa, como 'valor esperado' o, también, 'esperanza matemática'. Por último, algunos índices no tienen el privilegio de disfrutar de una doble asignación simbólica en función de que se traten de parámetros o de estadísticos, sea el caso de la mediana (*Md*) o el coeficiente de variación (*CV*), entre otros muchos.

• La aplicación sobre la distribución de probabilidad de una variable aleatoria de los índices de tendencia central, dispersión, etc. implica ciertas adaptaciones en las fórmulas presentadas para los mismos en los temas previos. A título ilustrativo, se muestran a continuación las de la media (o valor esperado) y la varianza para el caso en que la variable (*X*) sea ordinal o cuantitativa discreta:

$$E(X) = \mu_X = \sum X_i \cdot P(X_i)$$
$$\sigma_X^2 = \sum (X_i - \mu_X)^2 \cdot P(X_i)$$

Si las variables son cuantitativas continuas, las fórmulas se complican bastante más, pues interviene el cálculo integral en su aplicación. Pueden consultarse las mismas en cualquier libro de estadística avanzada.

Ejercicio 4: Obtener a partir del ejemplo presentado antes de la distribución de probabilidad de la variable "Nº de contratos laborales en los 2 últimos años":

a) ¿Cuál es la probabilidad de que una persona seleccionada al azar de la población anterior haya tenido 3 contratos en los 2 últimos años?
b) ¿Y de que haya tenido más de 2 contratos?
c) Obtener la mediana y la moda de la variable.
d) Obtener el valor esperado (media) y la varianza de la variable.

Ejercicio 5: Sea el caso que contamos con datos de dos variables, "Nº accidentes laborales en el último año" (*X*) y "Tipo de contrato" (*Y*) [Fijo; Temporal], en la población de trabajadores del sector de la construcción de Gandía (n = 1000). La distribución conjunta de frecuencias absolutas de ambas variables aleatorias se muestra en la siguiente tabla de contingencia:

Y_j \ X_i	0	1	2	3	
Fijo	250	90	50	40	430
Temporal	150	160	160	100	570
	400	250	210	140	1000

a) Obtener la distribución de probabilidad (función de probabilidad) de la variable *X* [$P(X_i)$]
b) Obtener la función de distribución de *X* [$P_a(X_i)$]
c) Obtener la distribución de probabilidad conjunta de ambas variables [$P(X_i, Y_j)$]
d) ¿Cuál es la probabilidad de que un trabajador (extraído al azar de dicha población)…
    1. haya tenido 1 o más accidentes? [$P(X \geq 1)$]
    2. tenga contrato fijo y haya tenido 0 accidentes? [P(Fijo$\cap$0)]
    3. haya tenido 2 o 3 accidentes? [$P(2 \cup 3)$]
    4. tenga un contrato fijo? [P(Fijo)]
    5. haya tenido 0 accidentes, sabiendo que tiene un contrato fijo? [$P(0/Fijo)$]
e) Obtener la moda, la mediana, la esperanza matemática y la varianza de la variable *X*.

• La siguiente tabla resume algunos de los conceptos planteados hasta ahora, diferenciados en función de que hagan referencia a una muestra o a una población:

MUESTRA	*POBLACIÓN*
1. Variable estadística	1. Variable aleatoria
2. Frecuencia relativa [p_i]	2. Probabilidad [$P(X_i)$]
3. Distribución de frecuencias relativas	3. Distribución de probabilidad → Dos tipos: Función de probabilidad Función de densidad de probabilidad
4. Frecuencia relativa acumulada [p_a]	4. Probabilidad acumulada [$P_a(X_i)$]
5. Distribución de frecuencias relativas acumuladas	5. Función de distribución
6. Estadístico	6. Parámetro

Referencias

Barón-López, J. (2005). Bioestadística: métodos y aplicaciones. Apuntes y material disponible en http://www.bioestadistica.uma.es/baron/apuntes/

Botella, J., León, O. G., San Martín, R. y Barriopedro, M. I. (2001). *Análisis de datos en psicología I: teoría y ejercicios.* Madrid: Pirámide.