Probabilidad - Clase 6
Teorema de Bernoulli y Grandes desvíos

Ernesto Mordecki

Facultad de Ciencias, Universidad de la República. Montevideo, Uruguay

Curso de la Licenciatura en Matemática - 2020

Contenidos

Figura: No Rogelio, yo ya no creo en las estadísticas, desde que me enteré que 3 de cada 10 son falsas

Esquema de Bernoulli

- El espacio de probabilidad es

$$\Omega_n = \{\omega = (\omega^{(1)}, \dots, \omega^{(n)}) : \text{cada } \omega^{(i)} \text{ es éxito } \omega_1, \text{ o fracaso } \omega_2\ (i = 1, \dots, n).\}$$

- La σ-álgebra es $\mathcal{P}(\Omega_n)$
- la probabilidad de cada suceso elemental

$$\mathbf{P}(\omega) = p^k q^{n-k}.$$

donde k es la cantidad de éxitos de ω

Distribución binomial

- $\mu(\omega)$ es cantidad de éxitos en $\omega \quad (0 \leq \mu(\omega) \leq m.)$
- Tenemos un teorema que prueba que

$$P_n(m) = \mathbf{P}(\mu = m) = \binom{n}{m} p^m q^{n-m}.$$

para $m = 0, 1, \ldots, n$.

Estadística: estimación de p

Problema: Observamos los resultados de un esquema de Bernoulli con n experimentos. Desconocemos $p \in (0, 1)$. Nos proponemos *estimar* el parámetro p. Mas precisamente:

- Sabemos que observamos un esquema de Bernoulli
- Observamos una realización ω. En particular sabemos $\mu(\omega)$
- Desconocemos $p \in (0, 1)$
- ¿Cómo podemos calcular (o al menos aproximar) p?
- una propuesta intuitiva es tomar:

Estimación de máxima verosimiltud

$$\hat{p} \sim \frac{\mu(\omega)}{n}$$

$\boxed{\text{¿Podemos dar un fundamento a esta intuición?}}$

La propuesta es: Calculemos la probabilidad de la observación ω y elijamos el p que haga máxima esa probabilidad. Tenemos

- $\mathbf{P}(\omega, p) = p^k q^{n-k}$
- Ahora conocemos k pero desconocemos p (por eso escribimos la probabilidad como función de p).
- Resolvemos el problema

$$\max \mathbf{P}(\omega, p) \text{ para } p \in (0, 1)$$

Hagamos las cuentas

- El máximo de $\mathbf{P}(\omega, p)$ es el mismo que el de $\log \mathbf{P}(\omega, p)$. Tenemos entonces

$$\ell(p) = \log \mathbf{P}(\omega, p) = \log p^k q^{n-k} = k \log p + (n-k) \log q$$

- Derivamos

$$\frac{\partial \ell(p)}{\partial p} = \frac{k}{p} - \frac{n-k}{1-p} = 0.$$

- Operando: $k(1-p) - p(n-k) = k - kp - pn + pk = 0$
- De aquí:

$$\hat{p} = \frac{k}{n} \text{ es el } \textit{estimador de máxima verosimilitud}.$$

Repasando

- Tenemos una estrategia de estimación, la *máxima verosimilitud* (máxima “creencia”)
- Calculamos la probabilidad de obtener la *muestra* que observamos, como función del parámetro (p en este caso) a estimar
- Calculamos el máximo de esa probabilidad como función de p
- Obtenemos el estimador de máxima verosimilitud $\hat{p}$.
- El sombrero $\hat{\theta}$ se usa para indicar que se está *estimando* un parámetro θ desconocido.

Conclusión

- En probabilidades tenemos un espacio
$$(\Omega, \mathcal{A}, \mathbf{P})$$
con **P** conocida.
- A partir de esto calculamos probabilidades de sucesos que nos interesan. Después ocurre el azar (experimento)
- En estadística los experimentos se realizan primero, surgen *datos*, el ω, pero se desconoce **P**
- Con esos datos se *estima* la probabilidad **P**

Ley de los grandes números de Bernoulli

- Supongamos que repetimos un experimento con dos resultados posibles que denominamos éxito y fracaso
- Designemos mediante μ la cantidad de éxitos.
- Esto incluye el caso particular de tirar una moneda n veces y contar la cantidad de veces que obtenemos cara.

- Queremos estudiar, para valores grandes de *n*, el comportamiento del cociente μ/n, es decir, de la frecuencia de éxitos.
- Este comportamiento viene dado por el siguiente resultado, obtenido por J. Bernoulli, aparecido en 1713.
- Se fundamenta la intuición, que dice que la frecuencia asintótica de los éxitos es *p*. En el caso de tirar una moneda, esta intuición dice que, “aproximadamente, la mitad de las veces aparce cara”.

Teorema de Bernoulli. Ley de los grandes números

Teorema

Dados $\varepsilon > 0$, $\eta > 0$, números arbitrariamente pequeños, existe un número natural n_0 tal que, para todo $n \geq n_0$,

$$\mathbf{P}\left(\left|\frac{\mu}{n} - p\right| < \varepsilon\right) \geq 1 - \eta.$$

En palabras,

- dado $\varepsilon > 0$ la probabilidad de que la diferencia entre la frecuencia observada $\frac{\mu}{n}$ y p sea menor que ε es arbitrariamente cercana a 1 cuando n es suficientemente grande.

Para demostrar este teorema, recordando que

$$\sum_{k=0}^{n} P_n(k) = \sum_{k=0}^{n} C_k^n p^k q^{n-k} = 1, \tag{1}$$

obtenemos primero dos fórmulas auxiliares

Lema

Valen las siguientes identidades:

$$\sum_{k=0}^{n} k P_n(k) = np, \tag{2}$$

$$\sum_{k=0}^{n} k^2 P_n(k) = n^2 p^2 + npq. \tag{3}$$

Demostración del Teorema

Comencemos observando que los dos sucesos $\left|\frac{\mu}{n} - p\right| < \varepsilon$, $\left|\frac{\mu}{n} - p\right| \geq \varepsilon$ son contrarios, por lo que

$$\mathbf{P}\left(\left|\frac{\mu}{n} - p\right| < \varepsilon\right) = 1 - \mathbf{P}\left(\left|\frac{\mu}{n} - p\right| \geq \varepsilon\right)$$

Según la regla de la suma de probabilidades, tenemos

$$\mathbf{P}\left(\left|\frac{\mu}{n} - p\right| \geq \varepsilon\right) = \sum P_n(k)$$

donde la suma abarca todos los k tales que $\left|\frac{k}{n} - p\right| \geq \varepsilon$. Para estos valores de k se verifica también

$$\frac{\left(\frac{k}{n} - p\right)^2}{\varepsilon^2} \geq 1.$$

Basados en esta última desigualdad, obtenemos

$$\mathbf{P}\left(\left|\frac{\mu}{n} - p\right| \geq \varepsilon\right) \leq \sum \frac{\left(\frac{k}{n} - p\right)^2}{\varepsilon^2} P_n(k)$$

donde la suma se extiende a los mismos valores de k que antes. Es claro que al sumar en todos los valores de k desde 0 a n, obtenemos un resultado mayor.

En consecuencia

$$\mathbf{P}\left(\left|\frac{\mu}{n}-p\right|\geq\varepsilon\right)\leq\sum_{k=0}^{n}\frac{\left(\frac{k}{n}-p\right)^{2}}{\varepsilon^{2}}P_{n}(k)=\frac{1}{n^{2}\varepsilon^{2}}\sum_{k=0}^{n}(k-np)^{2}P_{n}(k)$$

$$=\frac{1}{n^{2}\varepsilon^{2}}\left[\sum_{k=0}^{n}k^{2}P_{n}(k)-2np\sum_{k=0}^{n}kP_{n}(k)+n^{2}p^{2}\sum_{k=0}^{n}P_{n}(k)\right]$$

$$=\frac{1}{n^{2}\varepsilon^{2}}\left[n^{2}p^{2}+npq-2n^{2}p^{2}+n^{2}p^{2}\right]=\frac{pq}{n\varepsilon^{2}}$$

En los últimos cálculos utilizamos las fórmulas (1), (2) y (3).
Concluyendo

$$\mathbf{P}\left(\left|\frac{\mu}{n}-p\right|<\varepsilon\right)=1-\mathbf{P}\left(\left|\frac{\mu}{n}-p\right|\geq\varepsilon\right)\geq 1-\frac{pq}{n\varepsilon^{2}}. \qquad (4)$$

Esta desigualdad demuestra el teorema.

Aplicación

Nos interesa calcular la probabilidad de que entre 400 niños que nacen, la la cantidad de varones se desvie del valor que se espera, que es 200, en no más de 20. **Solución** Aplicamos el ejemplo anterior, suponiendo la probabilidad de nacimiento de un varón igual a $1/2$, y μ cuenta la cantidad de varones que nacen, con un total de $n = 400$ nacimientos. Tenemos que calcular

$$\mathbf{P} = \mathbf{P}\left(|\mu - 200| < 20\right) = \mathbf{P}\left(\left|\frac{\mu}{n} - \frac{1}{2}\right| < \frac{20}{n}\right).$$

Dado que el cálculo numérico de esta probabilidad es díficil, utilizamos la cota obtenida en el teorema. Tenemos $\varepsilon = 20/400 = 1/20$, y obtenemos

$$\mathbf{P} \geq 1 - \frac{1}{4n^2\varepsilon^2} = 1 - \frac{1}{4 \times 400 \times (1/20)^2} = 3/4.$$

Un cálculo aproximado, pero más preciso, establece que $\mathbf{P} \sim 0,95$.

Consecuencias en estadística del Teorema de Bernoulli

- Consideramos el problema de estimación de p a partir de una *muestra* que observamos: ω.
- El estimador de máxima verosimilitud de p es, según vimos
$$\hat{p} = \frac{k}{n}$$
- El teorema de Bernoulli establece que
$$\mathbf{P}\left(\left|\hat{p} - p\right| > \varepsilon\right) \leq \frac{pq}{n\varepsilon^2} \to_n 0.$$

- También decimos que

$$\hat{p} \text{ converge en probabilidad a } p,$$

cuando ocurre el resultado anterior, y se escribe

$$\hat{p} \xrightarrow{\mathbf{P}} p, \quad \text{cuando } n \to \infty$$

- En estadística se dice que

$\hat{p}$ es es un *estimador insesgado* de p.

Desigualdad de grandes desvíos para el Esquema de Bernoulli

- Consideremos un esquema de Bernoulli con n experimentos, probabilidad de éxito p, $0 < p < 1$, y $q = 1 - p$.
- Esperamos que $\mu/n \sim p$.
- Nos interesa encontrar una **cota** para la probabilidad de que la frecuencia μ/n se desvíe de el valor que esperamos p en una cierta cantidad.
- Elegimos $\delta = p + \varepsilon > p$ y acotamos

$$\mathbf{P}\left(\frac{\mu}{n} \geq \delta\right).$$

Cota superior de la probabilidad en grandes desvíos

Veamos Para acotar elegimos $\theta > 0$ auxiliar. Haremos una cuenta parecida a la demostración del Teorema de Bernoulli (el truco del **1**):

$$1 \leq \frac{e^{\theta k}}{e^{\theta \delta n}}$$

para $k \geq n\delta$**. Tenemos:**

$$\begin{aligned}
\mathbf{P}\left(\frac{\mu}{n} \geq \delta\right) &= \sum_{k \geq n\delta} \mathbf{P}(\mu = k) = \sum_{k \geq n\delta} 1 \binom{n}{k} p^k q^{n-k} \\
&\leq \sum_{k \geq n\delta} \frac{e^{\theta k}}{e^{\theta \delta n}} \binom{n}{k} p^k q^{n-k} \\
&\leq e^{-\theta \delta n} \sum_{k=1}^{n} e^{\theta k} \binom{n}{k} p^k q^{n-k} \\
&= e^{-\theta \delta n} \left(p e^{\theta} + q\right)^n \\
&= \exp\left(-n\left[\theta\delta - \log\left(p e^{\theta} + q\right)\right]\right).
\end{aligned}$$

Consideramos entonces la función (auxiliar también):

$$f(\theta) = \theta\delta - \log\left(pe^{\theta} + q\right),$$

Lo que obtuvimos es

$$\mathbf{P}\left(\frac{\mu}{n} \geq \delta\right) \leq \exp\left(-nf(\theta)\right).$$

y elegimos θ para maximizarla, lo que nos va a dar la mejor cota con esta metodología.

Derivamos

$$f'(\theta) = \delta - \frac{pe^{\theta}}{pe^{\theta} + q},$$

de donde el máximo se da en

$$\theta^* = \log(\delta q/(1-\delta)p),$$

y el valor máximo $f'(\theta^*)$ (después de algunas cuentas) es

$$H(\delta) = \delta \log \frac{\delta}{p} + (1-\delta) \log \frac{1-\delta}{q}.$$

Observemos que $H(p) = 0$, la cota nos da uno (es una cota no informativa). Pero para $\delta > p$ obtenemos una acotación útil

Cota de Grandes desvíos

Teorema. Para un esquema de Bernoulli, dado $\delta > p$ se tiene la *desigualdad de grandes desvíos*:

$$\mathbf{P}\left(\frac{\mu}{n} \geq \delta\right) \leq e^{-nH(\delta)},$$

En el caso simétrico, cuando $p = q = 1/2$, tenemos

$$H(\delta) = \delta \log \delta + (1-\delta)\log(1-\delta) + \log 2.$$

Observemos que de la desigualdad obtenida, resulta que

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbf{P}\left(\frac{\mu}{n} \geq \delta\right) \leq -H(\delta).$$

Cota inferior del intervalo

- Queremos acotar la frecuencia por valores menores:

$$\gamma = p - \varepsilon \leq \frac{\mu}{n}$$

- Así tenemos acotada la probabilidad de que

$$\left|\frac{\mu}{n} - p\right| \geq \varepsilon.$$

- De es forma obtenemos una acotación del mismo suceso que en el teorema de Bernoulli
- La cota era

$$\mathbf{P}\left(\left|\frac{\mu}{n} - p\right| > \varepsilon\right) \leq \frac{pq}{n\varepsilon^2}.$$

- Para $\gamma < p$, queremos acotar

$$\mathbf{P}\left(\frac{\mu}{n} \leq \gamma\right).$$

- **Estrategia:** ¡convertir los fracasos en éxitos!,
- Contamos

$$\tilde{\mu} = n - \mu.$$

- tenemos un esquema con probabilidad de éxito q.

Tenemos

$$\mathbf{P}\left(\frac{\mu}{n} \leq \gamma\right) = \mathbf{P}\left(1 - \frac{\mu}{n} \geq 1 - \gamma\right)$$
$$= \mathbf{P}\left(\frac{n-\mu}{n} \geq 1 - \gamma\right) = \mathbf{P}\left(\frac{\tilde{\mu}}{n} \geq 1 - \gamma\right) \leq e^{-n\tilde{H}(1-\gamma)}.$$

donde $\tilde{H}$ se obtiene de H cambiando p por q, es decir

$$\tilde{H}(\delta) = \delta \log \frac{\delta}{q} + (1-\delta) \log \frac{1-\delta}{p} = H(1-\delta),$$

y la cota es

$$\mathbf{P}\left(\frac{\mu}{n} \leq \gamma\right) \leq \exp\left(-nH(\gamma)\right).$$

Teorema de Grandes Desvíos

Teorema. Para un esquema de Bernoulli con probabilidad p de éxito, se tiene

$$\begin{aligned}\mathbf{P}\left(\left|\frac{\mu}{n}-p\right|\geq\varepsilon\right)&=\mathbf{P}\left(\frac{\mu}{n}\geq p+\varepsilon\right)+\mathbf{P}\left(\frac{\mu}{n}\leq p-\varepsilon\right)\\&\leq\exp\left(-nH(p+\varepsilon)\right)+\exp\left(-nH(p-\varepsilon)\right).\end{aligned}$$

En el caso simétrico, obtenemos

$$\mathbf{P}\left(\left|\frac{\mu}{n}-\frac{1}{2}\right|\geq\varepsilon\right)\leq 2\exp\left(-nH(\varepsilon)\right)$$

donde

$$H(\varepsilon)=(1/2-\varepsilon)\log(1-2\varepsilon)+(1/2+\varepsilon)\log(1+2\varepsilon)$$

Veamos algunas comparaciones numéricas: En el ejemplo de los nacimientos, tenemos $n = 400$, $\varepsilon = 1/20$, y obtenemos

$$\mathbf{P}\left(\left|\frac{\mu}{n} - \frac{1}{2}\right| \geq \varepsilon\right) \leq 2\exp\left(-400H(1/20)\right))$$

Aquí

$$H(1/20) = (9/20)\log(18/20) + (11/20)\log(22/20).$$

La cuenta da

$$\mathbf{P}\left(\left|\frac{\mu}{n} - \frac{1}{2}\right| \geq \varepsilon\right) \leq 0.27.$$

La cota de Chevishev nos da 0.25.

- Si $n = 800$, tenemos una cota de grandes desvíos de 0.036,
- la cota de Chebishev da 0.125.

Cota inferior de la probabilidad en grandes desvíos

Obtuvimos

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbf{P}\left(\frac{\mu}{n} \geq \delta\right) \leq -H(\delta).$$

Veamos que este límite superior es efectivamente un límite.
Acotamos inferiormente con $k = [n\delta] + 1$:

$$\mathbf{P}\left(\frac{\mu}{n} \geq \delta\right) \geq \mathbf{P}\left(\mu = k\right) = C_k^n p^k (1-p)^{n-k}.$$

de donde

$$\frac{1}{n} \log \mathbf{P}\left(\frac{\mu}{n} \geq \delta\right) \geq \frac{1}{n} \log\left(C_k^n p^k (1-p)^{n-k}\right).$$

Recordamos la fórmula de Stirling para el factorial:

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$$

La cuenta es:

$$\begin{aligned}
C_k^n p^k(1-p)^{n-k} &= \frac{n!}{k!(n-k)!}p^k(1-p)^{n-k} \\
&\sim \frac{\sqrt{2\pi n}}{\sqrt{2\pi k}\sqrt{2\pi(n-k)}}\left(\frac{n}{e}\right)^n\left(\frac{e}{k}\right)^k\left(\frac{e}{n-k}\right)^{n-k} \\
&= \sqrt{\frac{n}{2\pi k(n-k)}}\left(\frac{n}{k}\right)^k\left(\frac{n}{n-k}\right)^{n-k} p^k(1-p)^{n-k} \\
&= \sqrt{\frac{n}{2\pi k(n-k)}}\left(\frac{np}{k}\right)^k\left(\frac{n(1-p)}{n-k}\right)^{n-k}
\end{aligned}$$

Ahora tomamos logaritmo.

Primero vemos que

$$\frac{k}{n} = \frac{[n\delta]+1}{n} \to \delta.$$

La cuenta es:

$$\begin{aligned}
&\frac{1}{n}\log\left(C_k^n p^k (1-p)^{n-k}\right) \\
&\qquad \sim \frac{1}{n}\log\left(\sqrt{\frac{n}{2\pi k(n-k)}}\left(\frac{np}{k}\right)^k\left(\frac{n(1-p)}{n-k}\right)^{n-k}\right) \\
&= \frac{1}{2n}\log\left(\frac{n}{2\pi k(n-k)}\right) - \frac{k}{n}\log\left(\frac{k/n}{p}\right) - \frac{n-k}{n}\log\left(\frac{1-k/n}{1-p}\right) \\
&\qquad \to -\delta\log\left(\frac{\delta}{p}\right) - (1-\delta)\log\left(\frac{1-\delta}{1-p}\right) = -H(\delta),
\end{aligned}$$

En vista de (5), obtenemos la velocidad de los grandes desvíos en el esquema de Bernoulli:

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbf{P}\left(\frac{\mu}{n} \geq \delta\right) = -H(\delta). \tag{5}$$

Comparación de las cotas de Bernuolli y de grandes desvíos

La idea central de las acotaciones es acotar el error de la estimación

$$\frac{\mu}{n} \sim p.$$

Para eso se asume un error ε tolerable y se estudia la probabilidad de cometer un error mayor, es decir la probabilidad

$$\mathbf{P}\left(\left|\frac{\mu}{n} - p\right| \geq \varepsilon\right).$$

La cota de Bernoulli fué obtenida en 1713, mientras que el teorema de grandes desvíos, en una forma más general que la presentada, fue obtenido por Cramér en 1936. En la Figura 38 vemos el error de ambas fórmulas como función de n para $\varepsilon = 0,05$.

Tenemos

$$H(1/2+0,05) = H(0,55) = 0,45\log 0,9+0,55\log 1,1 = 0.005.$$

Las cotas de Bernuolli y de grandes desvíos son entonces

$$B(n) = \frac{1}{4(0,05)^2 n} = \frac{100}{n}, \qquad GD(n) = 2\exp(-0.005n).$$