Generalidades suma de variables aleatorias

Sean X e Y dos variables aleatorias. Luego,

$$E[X+Y] = E[X] + E[Y],$$

$$\mathrm{Var}[X+Y] = \mathrm{Var}[X] + 2\mathrm{Cov}[X,Y] + \mathrm{Var}[Y].$$

Si $\mathrm{Cov}[X,Y] = 0$, por ejemplo en el caso en que sean dos variables aleatorias independientes,

$$\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y].$$

En el caso en que se trata de variables aleatorias independientes, también tenemos

$$f_{X+Y}(z) = \int_{-\infty}^{+\infty} f_X(x) f_Y(z-x)\,dx = \int_{-\infty}^{+\infty} f_X(z-y) f_Y(y)\,dy,$$

si se trata de dos variables continuas,

$$P(X+Y=z) = \sum_{x\in\mathcal{R}_X} P(X=x)P(Y=z-x) = \sum_{y\in\mathcal{R}_Y} P(X=z-y)P(Y=y),$$

si son discretas, donde $\mathcal{R}_X$ y $\mathcal{R}_Y$ son los recorridos de X y de Y, respectivamente.

Sean $\{X_k\}_{k=1}^n$ variables aleatorias independientes y llamemos

$$S_n = \sum_{k=1}^n X_k.$$

Luego,

$$E[S_n] = \sum_{k=1}^n E[X_k], \qquad \mathrm{Var}[S_n] = \sum_{k=1}^n \mathrm{Var}[X_k].$$

En el caso particular en que, además, las X_k estén idénticamente distribuidas,

$$E[S_n] = nE[X_1], \qquad \mathrm{Var}[S_n] = n\mathrm{Var}[X_1].$$

Algunas sumas especiales

1. **Suma de v.a. Bernoulli es binomial:** Sean $X_1, X_2, \cdots, X_n \sim \text{Bernoulli}(p)$ independientes. Luego, $\sum_{k=1}^{n} X_k \sim \text{Binomial}(n, p)$.

2. **Suma de v.a. binomiales es binomial:** Sean $N_1 \sim \text{Binomial}(n, p)$ y $N_2 \sim \text{Binomial}(m, p)$ independientes. Luego $N_1 + N_2 \sim \text{Binomial}(n + m, p)$. El ejemplo anterior es un caso especial.

3. **Suma de v.a. Poisson es Poisson:** Sean $N_1 \sim \text{Poisson}(\lambda_1)$ y $N_2 \sim \text{Poisson}(\lambda_2)$ independientes. Luego $N_1 + N_2 \sim \text{Poisson}(\lambda_1 + \lambda_2)$.

4. **Suma de v.a. normales es normal:** Sean $X_1 \sim \mathcal{N}(\mu_1, \sigma_1)$ y $X_2 \sim \mathcal{N}(\mu_2, \sigma_2)$ independientes. Luego $X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2})$.

5. **Suma de v.a. exponenciales es gama:** Sean $X_1, X_2, \cdots, X_n \sim \text{Expo}(\lambda)$ independientes. Luego, $\sum_{k=1}^{n} X_k \sim \Gamma(n, \lambda)$, donde

$$f_{\Gamma(n,\lambda)}(x) = \frac{(\lambda x)^{n-1}}{(n-1)!} \lambda e^{-\lambda x} \qquad \text{para } x > 0.$$

6. **Suma de los cuadrados de v.a. normales es chi(ji)-cuadrado:** Sean $X_1, X_2, \cdots, X_n \sim \mathcal{N}(0, 1)$ independientes. Luego, $\sum_{k=1}^{n} X_k^2 \sim \chi_n^2$ (chi(ji)-cuadrado con n grados de libertad), donde

$$f_{\chi_n^2}(x) = \left(\frac{x}{2}\right)^{\frac{n}{2}} \frac{e^{-\frac{x}{2}}}{x\Gamma\left(\frac{n}{2}\right)} \qquad \text{para } x > 0.$$

7. **Suma de v.a. chi-cuadrado es chi-cuadrado:** Sean $X_1 \sim \chi_n^2$ y $X_2 \sim \chi_m^2$ independientes. Luego $X_1 + X_2 \sim \chi_{n+m}^2$.

Distribuciones infinitamente divisibles

Se dice que una variable aleatoria X tiene una *distribución infinitamente divisible* si para todo $n \in \mathbb{N}$ puede ser escrita como suma de variables i.i.d.

Ejemplos:

1. Poisson.
2. Gama.
3. Normal.
4. χ^2.

Ejercicio 5 de la guía 7

Suponga que 30 instrumentos electrónicos $D_1, D_2, \cdots, D_{30}$ se usan de la siguiente manera: tan pronto como D_1 falla empieza a actuar D_2, cuando éste falla empieza a actuar D_3 y así sucesivamente. Si la variable aleatoria asociada a la duración de cada instrumento tiene distribución exponencial de parámetro 0.1 1/hora y T es el tiempo total de operación de los 30 instrumentos, ¿cuál es la probabilidad de que T exceda 310 horas? Se supone que los tiempos de operación son variables aleatorias independientes.

Respuesta:
Sabemos que $T \sim \Gamma(30, 0.1)$. El ejercicio pide calcular

$$P(T > 310) = 1 - P(T \leq 310). \tag{1}$$

Esto es fácil de hacer utilizando un programa de cálculo. En Octave, por ejemplo, se escribe:

```
>> 1-gamcdf(310,30,1/0.1)
ans =  0.40465
```

Aquí `gamcdf(x,a,b)` es una función que calcula la función de distribución de probabilidades acumulada de una variable aleatoria $\Gamma(a, 1/b)$.

Otra forma de calcular el mismo valor, es usando la relación con una variable aleatoria Poisson. Sea $N \sim \text{Poisson}(310 \times 0.1)$. De lo que hemos estudiado del proceso de Poisson, sabemos que

$$P(T > 310) = P(N < 30) = \sum_{k=0}^{29} \frac{31^k}{k!} e^{-31}. \tag{2}$$

Este valor se puede calcular usando un programa matemático. En Octave:

```
>> poisscdf(29,31)
ans =  0.40465
```

`poisscdf(x,a)` es una función que calcula la función de distribución de probabilidades acumulada de una variable aleatoria Poisson(a).

Desigualdades de Markov y de Chebychev

Sea X una variable aleatoria cualquiera. Luego, la desigualdad de Markov nos dice que para todo $\varepsilon > 0$

$$P\big(|X| \geq \varepsilon\big) \leq \frac{E\left[|X|\right]}{\varepsilon}.$$

La desigualdad de Chebychev es una extensión de la de Markov: para todo $\varepsilon > 0$

$$P\big(|X - E\left[X\right]| \geq \varepsilon\big) \leq \frac{\mathrm{Var}\left[X\right]}{\varepsilon^2}.$$

Nada se asume respecto de las variables aleatorias en estas desigualdades.
La desigualdad de Chebychev se puede usar, por ej., para la suma S_n variables aleatorias i.i.d. (independientes e idénticamente distribuidas) $\{X_k\}_{k=1}^n$: para todo $\varepsilon > 0$

$$P\left(\left|\frac{S_n}{n} - E\left[X_1\right]\right| \geq \varepsilon\right) \leq \frac{\mathrm{Var}\left[X_1\right]}{n\varepsilon^2}.$$

Ejercicio 5 de la guía 7

Suponga que 30 instrumentos electrónicos $D_1, D_2, \cdots, D_{30}$ se usan de la siguiente manera: tan pronto como D_1 falla empieza a actuar D_2, cuando éste falla empieza a actuar D_3 y así sucesivamente. Si la variable aleatoria asociada a la duración de cada instrumento tiene distribución exponencial de parámetro 0.1 1/hora y T es el tiempo total de operación de los 30 instrumentos, ¿cuál es la probabilidad de que T exceda 310 horas? Se supone que los tiempos de operación son variables aleatorias independientes.

Respuesta:
Sabemos que $E\left[T\right] = 30 \times 1/0.1 = 300$. Observando que T sólo puede tomar valores no-negativos y utilizando la desigualdad de Markov,

$$P\left(T > 310\right) \leq \frac{300}{310} = \frac{30}{31} \approx 0.9677. \tag{3}$$

La cota es válida, pero no es muy ajustada (ver resultado más arriba).
También sabemos que $\mathrm{Var}\left[T\right] = 30 \times 1/0.1^2 = 3000$. Utilizando la desigualdad de Chebychev,

$$P\left(T > 310\right) = P\left(T - 300 > 10\right) \leq P\left(|T - 300| > 10\right) \leq \frac{3000}{10^2} = 30 > 1. \tag{4}$$

Claramente la cota es válida, pero no es para nada ajustada.
Las desigualdades de Markov y Chebychev no usan casi ninguna información sobre la variable aleatoria, por lo que no proveen cotas muy ajustadas.

Ley de los grandes números

Sean $\{X_k\}_{k=1}^n$ i.i.d. con $\mu = E[X_1]$ y $\sigma^2 = \text{Var}[X_1]$. Luego, la *ley débil de los grandes números* dice que para todo $\varepsilon > 0$

$$\lim_{n\to\infty} P\left(\left|\bar{X}_n - \mu\right| \geq \varepsilon\right) = 0,$$

donde

$$\bar{X}_n = \frac{S_n}{n} = \frac{1}{n}\sum_{k=1}^{n} X_k.$$

Claramente, esta ley es una consecuencia de la desigualdad de Chebychev. Intuitivamente, dice que la probabilidad de que el promedio y el valor medio difieran en algo se hace cada vez más pequeña a medida que se promedian más datos.

La *ley fuerte de los grandes números* afirma que

$$P\left(\lim_{n\to\infty} \bar{X}_n = \mu\right) = 1.$$

Se dice que

$$\lim_{n\to\infty} \bar{X}_n = \mu \qquad \text{con probabilidad } 1.$$

La diferencia entre ambas "leyes" es sutil, pero importante. La ley fuerte dice que la probabilidad de que el promedio difiera del valor medio es nula.

Sea $A \subseteq \Omega$ un evento, posible resultado de un experimento, y sea $p = P(A)$. Supongamos que el experimento se repite varias veces en forma independiente y definamos

$$\mathbf{1}_k(A) = \begin{cases} 1 & \text{si ocurrió } A \text{ en el } k\text{-ésimo experimento} \\ 0 & \text{en caso contrario.} \end{cases}$$

Es fácil ver que las variables aleatorias $\mathbf{1}_k(A)$ son i.i.d. con $E[\mathbf{1}_k(A)] = p$ y $\text{Var}[\mathbf{1}_k(A)] = p(1-p)$. Definamos la frecuencia relativa de ocurrencia del evento A en los primeros n experimentos como

$$\hat{P}_n = \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_k(A).$$

Luego, la ley fuerte de los grandes números nos dice que

$$\lim_{n\to\infty} \hat{P}_n = p \qquad \text{con probabilidad } 1.$$

Esta es la base de la interpretación frecuentista de la probabilidad: se pueden determinar las probabilidades mediante una larga repetición de experimentos.

Teorema Central del Límite

Sean $\{X_k\}_{k=1}^n$ i.i.d. con $\mu = E[X_1]$ y $\sigma^2 = \mathrm{Var}[X_1]$. Definamos

$$Z_n = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}.$$

Luego, el *teorema central del límite* dice que

$$\lim_{n\to\infty} F_{Z_n}(z) = \lim_{n\to\infty} P(Z_n \leq z) = \Phi(z),$$

donde $\Phi(z)$ es la función de distribución (acumulada) de probabilidad de una variable aleatoria normal estándar (con media cero y desvío 1).

Intuitivamente, el teorema nos dice que, a medida que n se hace más grande, Z_n se parece más a una variable $Z \sim \mathcal{N}(0,1)$. Por ello, para n grande (> 20) valen las siguientes aproximaciones:

$$\begin{aligned} F_{Z_n}(z) = P(Z_n \leq z) &\approx \Phi(z), \\ F_{\bar{X}_n}(x) = P(\bar{X}_n \leq x) &\approx \Phi\left(\frac{x-\mu}{\frac{\sigma}{\sqrt{n}}}\right), \\ F_{S_n}(s) = P(S_n \leq s) &\approx \Phi\left(\frac{s-n\mu}{\sqrt{n}\sigma}\right). \end{aligned}$$

Sea $A \subseteq \Omega$ un evento, posible resultado de un experimento, y sea $p = P(A)$. Supongamos que el experimento se repite varias veces en forma independiente. Sea $\hat{P}_n$ definida como más arriba. Luego, por el teorema central del límite tenemos que, para n muy grande

$$F_{\hat{P}_n}(q) = P\left(\hat{P}_n \leq q\right) \approx \Phi\left(\frac{q-p}{\sqrt{\frac{p(1-p)}{n}}}\right).$$

Corrección por continuidad

El teorema central del límite es válido tanto para variables aleatorias continuas como para variables aleatorias discretas. Sin embargo, para n grande (pero finito), estamos aproximando la distribución de una variable aleatoria discreta por la de una variable aleatoria continua. Para que la aproximación sea correcta, es necesario "ajustar" la distribución continua de la normal a la distribución discreta.
Supongamos, por ejemplo, que $\mathcal{R}_X = \{0, 1, 2, \cdots\}$ es el recorrido de las variables aleatorias $\{X_k\}_{k=1}^n$ i.i.d. con $\mu = E[X_1]$ y $\sigma^2 = \mathrm{Var}[X_1]$. Luego, $\mathcal{R}_{S_n} = \{0, 1, 2, \cdots\}$ y

$$\begin{aligned} P(S_n = s) = P\left(s - \frac{1}{2} < S_n \leq s + \frac{1}{2}\right) &\approx \Phi\left(\frac{s+\frac{1}{2}-n\mu}{\sqrt{n}\sigma}\right) - \Phi\left(\frac{s-\frac{1}{2}-n\mu}{\sqrt{n}\sigma}\right), \\ P(\bar{X}_n = x) = P\left(x - \frac{1}{2n} < \bar{X}_n \leq x + \frac{1}{2n}\right) &\approx \Phi\left(\frac{x+\frac{1}{2n}-\mu}{\frac{\sigma}{\sqrt{n}}}\right) - \Phi\left(\frac{x-\frac{1}{2n}-\mu}{\frac{\sigma}{\sqrt{n}}}\right), \end{aligned}$$

para todo $s \in \mathcal{R}_{S_n}$ y $x \in \mathcal{R}_{\bar{X}_n}$, respectivamente.

Ejercicio 5 de la guía 7

Suponga que 30 instrumentos electrónicos $D_1, D_2, \cdots, D_{30}$ se usan de la siguiente manera: tan pronto como D_1 falla empieza a actuar D_2, cuando éste falla empieza a actuar D_3 y así sucesivamente. Si la variable aleatoria asociada a la duración de cada instrumento tiene distribución exponencial de parámetro 0.1 1/hora y T es el tiempo total de operación de los 30 instrumentos, ¿cuál es la probabilidad de que T exceda 310 horas? Se supone que los tiempos de operación son variables aleatorias independientes.

Respuesta:
Sabemos que $E[T] = 30 \times 1/0.1 = 300$ y $\mathrm{Var}[T] = 30 \times 1/0.1^2 = 3000$. Dado que T es la suma de un número considerable de variables aleatorias i.i.d., podemos utilizar el teorema central del límite

$$P(T > 310) = 1 - P(T \leq 310) \approx 1 - \Phi\left(\frac{310 - 300}{\sqrt{3000}}\right) = 1 - \Phi\left(\frac{1}{\sqrt{30}}\right) \approx 0.4276. \tag{5}$$

Como se observa, el teorema central del límite proporciona una aproximación razonable al resultado correcto.

Ejercicio 1 de la guía 7

La probabilidad de dar en el blanco para un cañón es 0.8 (se supone constante y los disparos independientes). Calcular la probabilidad de que el número de blancos X satisfaga $70 < X < 90$ si se efectúan 100 disparos usando la distribución binomial y la aproximación normal a la distribución binomial.

Respuesta:
Dado que los disparos son independientes y la probabilidad de acertar constante, $X \sim \text{Binomial}(100, 0.8)$. Luego,

$$P(70 < X < 90) = P(71 \leq X \leq 89) = \sum_{k=71}^{89} \binom{100}{k} 0.8^k 0.2^{100-k} \approx 0.98305. \tag{6}$$

La cuenta se puede realizar utilizando un programa matemático. En Octave, por ej.,

```
>> binocdf(89,100,0.8)-binocdf(70,100,0.8)
ans =  0.98305
```

Aquí `binocdf(x,n,p)` es la función de distribución (acumulada) de una variable aleatoria Binomial(n, p). Usando Python, por ej.,

```
In [1] from scipy import stats as st

In [2] st.binom.cdf(89,100,0.8)-st.binom.cdf(70,100,0.8)
Out[2]: 0.98305464032321488
```

Sabemos que una variable aleatoria Binomial puede escribirse como la suma de variables aleatorias Bernoulli independientes. En el caso de X, ésta tiene la misma distribución que la suma de 100 v.a. i.i.d. Bernoulli(0.8). Dado que 100 es un número bastante grande, podemos utilizar el teorema central del límite para aproximar el valor de probabilidad requerido. Ahora bien, X es una v.a. discreta y aproximaremos una probabilidad referida a ella utilizando la distribución de una v.a. continua normal con media $\mu = 100 \times 0.8 = 80$ y desvío $\sigma = \sqrt{100 \times 0.8 \times 0.2} = 4$. Es por esto que resulta conveniente utilizar corrección por continuidad:

$$P(70 < X < 90) = P(70.5 < X \leq 89.5) \approx \Phi\left(\frac{89.5 - 80}{4}\right) - \Phi\left(\frac{70.5 - 80}{4}\right) \approx 0.98245. \tag{7}$$

La aproximación es exacta hasta el segundo decimal y tiene un error menor a 10^{-3}. Obsérvese que, al escribir la corrección por continuidad, se excluyó tanto al 70 como al 90 (se incluyeron en 71 y el 89).

También podríamos haber estimado la probabilidad solicitada utilizando la desigualdad de Chebychev. En este caso,

$$P(70 < X < 90) = P(|X - 80| \leq 9) = 1 - P(|X - E[X]| \geq 10) \geq 1 - \frac{16}{10^2} = 0.84. \tag{8}$$

Nuevamente, la cota provista por la desigualdad de Chebychev es correcta, pero poco ajustada.

Ejercicio 12 de la guía 7

Una fábrica produce determinados artículos de tal manera que una proporción p de ellos resultan defectuosos. Se inspeccionan n de tales artículos, y se determina la frecuencia relativa de defectuosos $\hat{P}_n$. a) Si $p = 0.07$, ¿cuál debería ser el tamaño n de manera tal que la probabilidad de que $\hat{P}_n$ difiera de p en menos de 0.01 sea al menos 0.98? Suponga válida la aproximación normal de la distribución binomial. b) Conteste la pregunta anterior si p se supone desconocida.

Respuesta:
El ejercicio pide encontrar n tal que

$$P\left(\left|\hat{P}_n - p\right| < 0.01\right) \geq 0.98. \tag{9}$$

Si los artículos a inspeccionar se sacan al azar con reposición o si se toman al azar de un conjunto de tamaño $N \gg n$, $X_n = n\hat{P}_n \sim \text{Binomial}(n, p)$, con $\mu_n = E[X_n] = np$ y $\sigma_n^2 = \text{Var}[X_n] = np(1-p)$. Luego, el ejercicio pide encontrar n tal que

$$P\left(\left|\hat{P}_n - p\right| < 0.01\right) = P(|X_n - np| < 0.01n) = \sum_{k=\lceil \mu_n - 0.01n \rceil}^{\lfloor \mu_n + 0.01n \rfloor} \binom{n}{k} p^k (1-p)^{n-k} \geq 0.98. \tag{10}$$

Claramente, este es un problema muy complicado. Si asumimos que n es grande, podemos aproxima la distribución de probabilidades de X_n por la de una variable aleatoria normal gracias al teorema central del límite. En este caso,

$$\begin{aligned} 0.98 \leq & P\left(\left|\hat{P}_n - p\right| < 0.01\right) \approx \\ & \Phi\left(\frac{\lfloor \mu_n + 0.01n \rfloor + 0.5 - \mu_n}{\sigma_n}\right) - \Phi\left(\frac{\lceil \mu_n - 0.01n \rceil - 0.5 - \mu_n}{\sigma_n}\right), \end{aligned} \tag{11}$$

donde hemos utilizado la corrección por continuidad. Sin embargo, σ_n crece con n como $\sqrt{n}$. Por lo tanto, la corrección por continuidad y el uso de la parte entera superior o inferior no hará mucha diferencia en las cuentas (es una corrección del orden de $1/\sqrt{n}$) cuando n es muy grande. Por lo tanto, podemos aproximar

$$0.98 \leq P\left(\left|\hat{P}_n - p\right| < 0.01\right) \approx \Phi\left(\frac{0.01n}{\sigma_n}\right) - \Phi\left(\frac{-0.01n}{\sigma_n}\right) = 2\Phi\left(\frac{0.01n}{\sigma_n}\right) - 1. \tag{12}$$

Basados en la aproximación, buscaremos n tal que

$$\Phi\left(\frac{0.01n}{\sigma_n}\right) \geq \frac{1+0.98}{2} = 0.99. \tag{13}$$

Luego,

$$\frac{0.01n}{\sigma_n} \geq z_{0.99} \Rightarrow 0.01n \geq z_{0.99}\sigma_n = z_{0.99}\sqrt{np(1-p)} \Rightarrow \sqrt{n} \geq \frac{z_{0.99}\sqrt{p(1-p)}}{0.01} \Rightarrow \tag{14}$$

$$n \geq 10000\, z_{0.99}^2 \left[p(1-p)\right]. \tag{15}$$

De una tabla podemos obtener $z_{0.99} = 2.3263$. Por lo tanto, para $p = 0.07$ tenemos

$$n \geq 3523.1 \Rightarrow n \geq 3524. \tag{16}$$

Se puede verificar que este resultado es muy cercano al correcto. Usando el hecho que $X_{3524} \sim$ Binomial$(3525, 0.07)$,

$$P\left(\left|\hat{P}_{3524} - 0.07\right| < 0.01\right) = P\left(211 < X_{3524} \leq 281\right) \approx 0.9792 \approx 0.98. \tag{17}$$

Para obtener el valor exacto, se puede utilizar, por ej., Octave

```
>> binocdf(281,3524,0.07)-binocdf(211,3524,0.07)
ans =  0.97916
```

o Python

```
In [1]: from scipy import stats as st

In [2]: st.binom.cdf(281,3524,0.07)-st.binom.cdf(211,3524,0.07)
Out[2]: 0.97915917917970574
```

Uno se puede preguntar: ¿todo n mayor a 3524 satisface aproximadamente el requerimiento? La Fig. 1 muestra que la respuesta es afirmativa para la mayor parte de los n entre 3524 y 10500. Más aún, sugiere que la respuesta es afirmativa para todos salvo unos pocos n mayores a 3524. El gráfico fue realizado utilizando las funciones de Octave.

Para comprender qué sucede en el caso en que p es desconocido, volvamos a la ecuación

$$n \geq 10000\, z_{0.99}^2 \left[p(1-p)\right]. \tag{18}$$

Dado que no conocemos p, debemos garantizar que n satisfaga esta desigualdad para todo valor de $p \in [0, 1]$, es decir,

$$n \geq 10000\, z_{0.99}^2 \max_{p\in[0,1]} \left[p(1-p)\right]. \tag{19}$$

Definamos $g(p) = p(1-p)$. Luego, $g(0) = g(1) = 0$ y

$$g'(p) = 1 - 2p, \qquad g''(p) = -2. \tag{20}$$

Es fácil ver que $g'(0.5) = 0$ y que $g(0.5) = 0.25$, por lo que el máximo de $g(p)$ se alcanza en $p = 0.5$ ($g''(p) < 0$). La Fig. 2 muestra $g(p)$. Por lo tanto, debemos pedir

$$n \geq 10000\, z_{0.99}^2 \, \frac{1}{4} = 13529.74 \Rightarrow n \geq 13530. \tag{21}$$

En este caso, n es mucho mayor que en el caso en que $p = 0.07$. Esto se debe a que n se ha calculado para el peor caso.

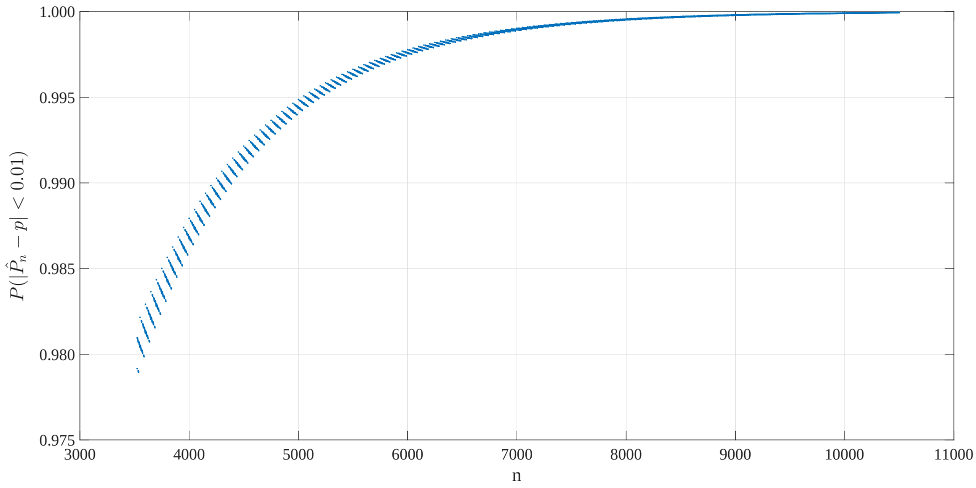

Figura 1: Respuesta al ejercicio 12 de la guía 7 cuando $p = 0.07$.

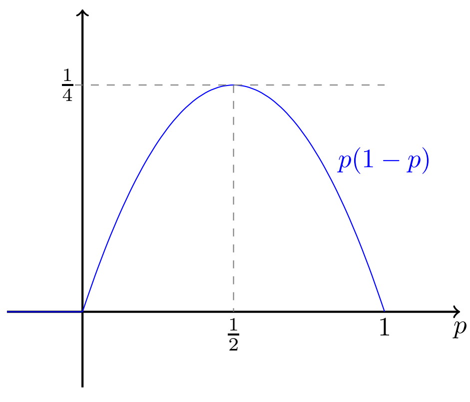

Figura 2: Varianza de una variable aleatoria Bernoulli: $p(1-p)$.