Probabilidad y Estadística (93.24)
Suma de variables aleatorias, distribución en el muestreo

Índice

1. Repaso de algunos conceptos

Generalidades suma de variables aleatorias

Sean X e Y dos variables aleatorias. Luego,

$$\mathrm{E}\left[X+Y\right]=\mathrm{E}\left[X\right]+\mathrm{E}\left[Y\right],$$

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

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

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

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

$$f_{X+Y}(z)=\int_{-\infty}^{+\infty}\mathrm{f}_X(x)\mathrm{f}_Y(z-x)\,dx=\int_{-\infty}^{+\infty}\mathrm{f}_X(z-y)\mathrm{f}_Y(y)\,dy,$$

si se trata de dos variables continuas,

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

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,

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

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

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

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. geométricas es binomial negativa:** Sean $X_1 \sim \text{Geo}(p)$ y $X_2 \sim \text{Geo}(p)$ independientes. Luego $X_1 + X_2 \sim \text{BinoNeg}(2, p)$.

6. **Suma de v.a. binomiales negativas es binomial negativa:** Sean $X_1 \sim \text{BinoNeg}(n, p)$ y $X_2 \sim \text{BinoNeg}(m, p)$ independientes. Luego $X_1 + X_2 \sim \text{BinoNeg}(n + m, p)$.

7. **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
$$\mathrm{f}_{\Gamma(n,\lambda)}(x) = \frac{(\lambda x)^{n-1}}{(n-1)!} \lambda e^{-\lambda x} \qquad \text{para } x > 0.$$

8. **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
$$\mathrm{f}_{\chi_n^2}(x) = \left(\frac{x}{2}\right)^{\frac{n}{2}} \frac{e^{-\frac{x}{2}}}{x\Gamma\left(\frac{n}{2}\right)} \qquad \text{para } x > 0.$$

9. **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 $X_1 + X_2 \sim \chi_{n+m}^2$.

10. **Suma de v.a. Cauchy:** Sean $X_1 \sim \text{Cauchy}(x_1, \gamma_1)$ y $X_2 \sim \text{Caucy}(x_2, \gamma_2)$ independientes. Luego $X_1 + X_2 \sim \text{Caucy}(x_1 + x_2, \gamma_1 + \gamma_2)$. Recordemos que
$$\mathrm{f}_{X_1}(x) = \frac{1}{\pi} \frac{\gamma}{\gamma^2 + (x - x_1)^2}.$$

Distribuciones infinitamente divisibles

Se dice que una variable aleatoria X tiene una *distribución infinitamente divisible* si puede ser escrita como suma de variables i.i.d.

Ejemplos:

1. Binomial negativa.
2. Poisson.
3. Gama.
4. Normal.
5. χ^2.
6. Cauchy.

Distribuciones estables

Sean X_1 y X_2 dos variables aleatorias independientes con la misma distribución que otra variable aleatoria X. Se dice que X es una v.a. estable si para cualquier $a_1, a_2 > 0$, existen $c > 0$ y d tales que $a_1X_1 + a_2X_2$ y $cX + d$ tienen la misma distribución. En palabras: una variable aleatoria es estable si una combinación lineal $(a_1X_1 + a_2X_2)$ de dos realizaciones independientes (X_1, X_2) de la misma tiene la misma distribución, salvo por un factor de escala (c) y de localización (d).

Es fácil ver que estable $\Rightarrow$ infinitamente divisible. Sin embargo, la dirección opuesta no es válida. Por ej., una variable aleatoria $X \sim \text{Poisson}(\lambda)$ es infinitamente divisible. Sabemos que $\mathcal{R}_X = \{0, 1, 2, \cdots\}$, por lo que $Y = cX + d$ tiene recorrido $\mathcal{R}_Y = \{d, c+d, 2c+d, \cdots\}$. Y podría ser Poisson sólo si $c = 1$ y $d = 0$.

Desigualdades de Markov y de Chebychev

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

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

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

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

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 $\varepsilon > 0$

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

Ley de los grandes números

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

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

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

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

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 $p = \mathrm{P}(A)$. 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 $\mathrm{E}[\mathbf{1}_k(A)] = p$ y $\mathrm{Var}[\mathbf{1}_k(A)] = p(1-p)$. 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 $\mu = \mathrm{E}[X_1]$ 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

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

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:

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

Sea $A \subseteq \Omega$ un evento, posible resultado de un experimento, y sea $p = \mathrm{P}(A)$. 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

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

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 $\mu = \mathrm{E}[X_1]$ y $\sigma^2 = \mathrm{Var}[X_1]$. Luego, $\mathcal{R}_{S_n} = \{0,1,2,\cdots\}$ y

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

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

2. Función de distribución de la variable aleatoria con distribución normal estándar

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
4.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

3. Fractiles de la distribución normal estándar

α	0.000	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009
0.50	0.0000	0.0025	0.0050	0.0075	0.0100	0.0125	0.0150	0.0175	0.0201	0.0226
0.51	0.0251	0.0276	0.0301	0.0326	0.0351	0.0376	0.0401	0.0426	0.0451	0.0476
0.52	0.0502	0.0527	0.0552	0.0577	0.0602	0.0627	0.0652	0.0677	0.0702	0.0728
0.53	0.0753	0.0778	0.0803	0.0828	0.0853	0.0878	0.0904	0.0929	0.0954	0.0979
0.54	0.1004	0.1030	0.1055	0.1080	0.1105	0.1130	0.1156	0.1181	0.1206	0.1231
0.55	0.1257	0.1282	0.1307	0.1332	0.1358	0.1383	0.1408	0.1434	0.1459	0.1484
0.56	0.1510	0.1535	0.1560	0.1586	0.1611	0.1637	0.1662	0.1687	0.1713	0.1738
0.57	0.1764	0.1789	0.1815	0.1840	0.1866	0.1891	0.1917	0.1942	0.1968	0.1993
0.58	0.2019	0.2045	0.2070	0.2096	0.2121	0.2147	0.2173	0.2198	0.2224	0.2250
0.59	0.2275	0.2301	0.2327	0.2353	0.2378	0.2404	0.2430	0.2456	0.2482	0.2508
0.60	0.2533	0.2559	0.2585	0.2611	0.2637	0.2663	0.2689	0.2715	0.2741	0.2767
0.61	0.2793	0.2819	0.2845	0.2871	0.2898	0.2924	0.2950	0.2976	0.3002	0.3029
0.62	0.3055	0.3081	0.3107	0.3134	0.3160	0.3186	0.3213	0.3239	0.3266	0.3292
0.63	0.3319	0.3345	0.3372	0.3398	0.3425	0.3451	0.3478	0.3505	0.3531	0.3558
0.64	0.3585	0.3611	0.3638	0.3665	0.3692	0.3719	0.3745	0.3772	0.3799	0.3826
0.65	0.3853	0.3880	0.3907	0.3934	0.3961	0.3989	0.4016	0.4043	0.4070	0.4097
0.66	0.4125	0.4152	0.4179	0.4207	0.4234	0.4261	0.4289	0.4316	0.4344	0.4372
0.67	0.4399	0.4427	0.4454	0.4482	0.4510	0.4538	0.4565	0.4593	0.4621	0.4649
0.68	0.4677	0.4705	0.4733	0.4761	0.4789	0.4817	0.4845	0.4874	0.4902	0.4930
0.69	0.4958	0.4987	0.5015	0.5044	0.5072	0.5101	0.5129	0.5158	0.5187	0.5215
0.70	0.5244	0.5273	0.5302	0.5330	0.5359	0.5388	0.5417	0.5446	0.5476	0.5505
0.71	0.5534	0.5563	0.5592	0.5622	0.5651	0.5681	0.5710	0.5740	0.5769	0.5799
0.72	0.5828	0.5858	0.5888	0.5918	0.5948	0.5978	0.6008	0.6038	0.6068	0.6098
0.73	0.6128	0.6158	0.6189	0.6219	0.6250	0.6280	0.6311	0.6341	0.6372	0.6403
0.74	0.6433	0.6464	0.6495	0.6526	0.6557	0.6588	0.6620	0.6651	0.6682	0.6713
0.75	0.6745	0.6776	0.6808	0.6840	0.6871	0.6903	0.6935	0.6967	0.6999	0.7031
0.76	0.7063	0.7095	0.7128	0.7160	0.7192	0.7225	0.7257	0.7290	0.7323	0.7356
0.77	0.7388	0.7421	0.7454	0.7488	0.7521	0.7554	0.7588	0.7621	0.7655	0.7688
0.78	0.7722	0.7756	0.7790	0.7824	0.7858	0.7892	0.7926	0.7961	0.7995	0.8030
0.79	0.8064	0.8099	0.8134	0.8169	0.8204	0.8239	0.8274	0.8310	0.8345	0.8381
0.80	0.8416	0.8452	0.8488	0.8524	0.8560	0.8596	0.8632	0.8669	0.8706	0.8742
0.81	0.8779	0.8816	0.8853	0.8890	0.8927	0.8965	0.9002	0.9040	0.9078	0.9116
0.82	0.9154	0.9192	0.9230	0.9269	0.9307	0.9346	0.9385	0.9424	0.9463	0.9502
0.83	0.9542	0.9581	0.9621	0.9661	0.9701	0.9741	0.9782	0.9822	0.9863	0.9904
0.84	0.9945	0.9986	1.0027	1.0069	1.0110	1.0152	1.0194	1.0237	1.0279	1.0322
0.85	1.0364	1.0407	1.0451	1.0494	1.0537	1.0581	1.0625	1.0669	1.0714	1.0758
0.86	1.0803	1.0848	1.0893	1.0939	1.0985	1.1031	1.1077	1.1123	1.1170	1.1217
0.87	1.1264	1.1311	1.1359	1.1407	1.1455	1.1503	1.1552	1.1601	1.1650	1.1700
0.88	1.1750	1.1800	1.1850	1.1901	1.1952	1.2004	1.2055	1.2107	1.2160	1.2212
0.89	1.2265	1.2319	1.2372	1.2426	1.2481	1.2536	1.2591	1.2646	1.2702	1.2759
0.90	1.2816	1.2873	1.2930	1.2988	1.3047	1.3106	1.3165	1.3225	1.3285	1.3346
0.91	1.3408	1.3469	1.3532	1.3595	1.3658	1.3722	1.3787	1.3852	1.3917	1.3984
0.92	1.4051	1.4118	1.4187	1.4255	1.4325	1.4395	1.4466	1.4538	1.4611	1.4684
0.93	1.4758	1.4833	1.4909	1.4985	1.5063	1.5141	1.5220	1.5301	1.5382	1.5464
0.94	1.5548	1.5632	1.5718	1.5805	1.5893	1.5982	1.6072	1.6164	1.6258	1.6352
0.95	1.6449	1.6546	1.6646	1.6747	1.6849	1.6954	1.7060	1.7169	1.7279	1.7392
0.96	1.7507	1.7624	1.7744	1.7866	1.7991	1.8119	1.8250	1.8384	1.8522	1.8663
0.97	1.8808	1.8957	1.9110	1.9268	1.9431	1.9600	1.9774	1.9954	2.0141	2.0335
0.98	2.0537	2.0748	2.0969	2.1201	2.1444	2.1701	2.1973	2.2262	2.2571	2.2904
0.99	2.3263	2.3656	2.4089	2.4573	2.5121	2.5758	2.6521	2.7478	2.8782	3.0902

4. Guía de ejercicios

1. Dentro de un ascensor, Francisco lee un cartel que dice: "Ocupación máxima: 4 personas. Peso máximo: 300 kg." Asumiendo que el peso medio de una persona tomada al azar sea de 75 kg, ¿puede dar una cota a la probabilidad de que 4 personas tomadas al azar superen en un 20 % el peso máximo permitido?

2. Repita el ejercicio anterior si se asume que el peso de una persona tomada al azar, X, puede modelarse como

   a) $X \sim \text{Unif}(0, 150)$.
   b) $X \sim \text{Exp}(1/75)$.
   c) $X \sim \mathcal{N}(75, 15)$.
   d) $X \sim \text{Poisson}(75)$ (en este caso, asumimos que el peso está redondeado a kilogramos).

3. 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.

   Resolución aquí.

4. Un sistema compuesto está formado por 100 componentes que funcionan independientemente. La probabilidad de que cualquier componente falle durante el período de operación es igual a 0.1. Para que el sistema completo funcione por lo menos deben funcionar 85 componentes. Calcular la probabilidad de que el sistema funcione correctamente usando la distribución binomial y la aproximación normal de la distribución binomial.

5. Una central telefónica A da servicio a 1800 usuarios de una central cercana B. Sería costoso y extravagante instalar 1800 líneas troncales de A a B. Es suficiente instalar un número N de líneas tan grande que, en condiciones ordinarias, solamente una de cada 100 llamadas en promedio no encuentre inmediatamente una línea troncal disponible. Supóngase que, en la hora más ocupada del día, cada usuario requiere una linea troncal de B durante un promedio de 2 minutos. En un momento fijo de la hora de máximo tráfico, puede suponerse la situación como un conjunto de 1800 ensayos independientes con una probabilidad $p = \frac{1}{30}$ (por lo de 2 min de cada 60 min) en que cada uno se requiere una línea. Determinar el número de líneas N a instalar entre A y B. Explicar claramente el planteo.

   Resolución aquí. Video con el ejercicio explicado.

6. Cierto proceso de fabricación de componentes electrónicos produce partes con un porcentaje de defectos del 5 %. Las partes son enviadas en cajas de 400 unidades. Las cajas con 25 o más partes defectuosas son devueltas.

   a) ¿Cuál es la probabilidad de que se devuelva una caja?
   b) En un día particular se enviaron 500 cajas. ¿Cuál es la probabilidad de que se devuelvan 80 o más cajas?
   c) Se introduce un nuevo proceso de fabricación que se supone reduce el procentaje de defectos. El objetivo es reducir la probabilidad de devolución de una caja a 0.01. ¿Cuál debe ser el porcentaje de defecto para que se alcance este objetivo.

7. Suponga que 30 instrumentos electrónicos $D_1, D_2, \ldots, D_{30}$ 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. La variable aleatoria T tiene *distribución Gamma*. Calcule la probabilidad pedida usando la distribución Gamma y la aproximación normal de esta suma de variables aleatorias independientes.

   Resolución aquí.

8. Al sumar números, una computadora aproxima cada número al entero más próximo. Supongamos que todos los errores de aproximación son independientes y distribuidos uniformemente en el intervalo (-0.5, 0.5). Si se suman 1500 números, ¿cuál es la probabilidad de que, en valor absoluto, el error total exceda 15?

9. Suponga que X_i, $i = 1, 2, \ldots, 50$, son variables aleatorias independientes que tienen cada una distribución de Poisson con parámetro $\lambda = 0.3$. Sea $S = X_1 + \cdots + X_{50}$.

   a) Usando el teorema central del límite calcular $\mathrm{P}(S \geq 18)$.

   b) Comparar la respuesta anterior con el valor exacto de esta probabilidad. Para calcular esta probabilidad considere que la suma de n variables aleatorias independientes con distribución de Poisson de parámetro λ tiene distribución de Poisson de parámetro $n\lambda$.

10. El número de accesos a un servidor web se supone (en cierto período del día) que es una variable aleatoria con distribución de Poisson con una media de 27 accesos por hora. Obtener la probabilidad de que haya al menos 90 accesos a lo largo de tres horas. Comparar los resultados obtenidos con la distribución de Poisson y con la aproximación normal.

11. Unos tambores etiquetados con un contenido de 30 litros son llenados con una solución proveniente de un gran tanque. El dispositivo de llenado de los tambores dispensa una cantidad por tambor que se supone una variable aleatoria con media de 30.01 litros y desvío estándar de 0.1 litros.

    a) ¿Cuál es la probabilidad de que la cantidad total de solución contenida en 50 tambores sea mayor a 1500 litros?

    b) Si la cantidad total de solución en el tanque es de 2401 litros, ¿cuál es la probabilidad de que puedan llenarse 80 tambores sin que se vacíe el tanque?

    c) ¿Qué cantidad de solución debe contener el tanque para que sea 0.9 la probabilidad de que puedan llenarse 100 tambores sin que el tanque se vacíe?

12. Los tiempos de servicio para los clientes que llegan a una caja de un supermercado pueden suponerse variables aleatorias independientes con un promedio de 1.5 minutos y una dispersión de 1 minuto.

    a) Calcular la probabilidad aproximada de que pueda atender a 100 clientes en menos de 2 horas de tiempo total de servicio de esta caja.

    b) Calcular el número de clientes n tal que la probabilidad de dar servicio a todos en menos de 2 horas sea aproximadamente 0.1.

13. Generar 10 números al azar, sumarlos, restarle 5 a la suma y dividir por $\sqrt{\frac{10}{12}}$ (¿un número mágico tal vez?). Repetir esa generación 2000 veces, por ejemplo. Hacer un histograma de los 2000 números obtenidos y comparar con la distribución normal estándar. Para los números aleatorios usar la función de generación de números aleatorios (RAND por ejemplo) de cualquier planilla de cálculo, programa utilitario matemático, calculadora o su lenguaje de programación preferido.

    Este código en *Octave* realiza la generación solicitada:

```
R = rand(10,2000);S = (sum(R)-5)/sqrt(10/12);
hist(S)
SS = sort(S);F =(1:2000)/2000;
plot(SS,F,SS,normcdf(SS))
D = max(abs(F-normcdf(SS)))
```

    En una ejecución de este código la distancia `D` fue `0.017934`. Aqui se compara la distribución empírica de los valores generados de `S` con la distribución normal estándar.

14. 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. En este caso recuerde que si $p \in (0, 1)$ entonces $p(1 - p) \leq 0.25$.

    Resolución aquí.

15. En un sistema están conectados 50 dispositivos. La probabilidad de que en el tiempo T un dispositivo cualquiera falle es 0.1. Utilizando la desigualdad de Tchebycheff estime la probabilidad de que la diferencia en valor absoluto entre el número de dispositivos que fallan y el promedio de los que fallan resulte: a) menor que 5 b) mayor que 5. Comparar el resultado con los que se obtienen si se utilizan el modelo binomial y su aproximación normal.

16. Se desea estimar la probabilidad p de ocurrencia de cierto suceso a partir de la frecuencia relativa de ocurrencia medida al realizar n repeticiones del experimento en que puede ocurrir el suceso de interés. Se pone como condición que la probabilidad de que el error (diferencia entre p y la frecuencia relativa) no supere el valor 0.05, sea mayor que 0.97. ¿Cuántas pruebas deben realizarse como mínimo? Comparar los resultados obtenidos a partir de usar:

    a) la desigualdad de Tchebycheff;

    b) la aproximación normal de la distribución binomial.

    Resolución aquí. Video con el ejercicio explicado.

17. El peso de las cajas transportadas por una compañía de transportes se distribuye de manera aproximadamente normal de media 20 kg y desviación estándar 3 Kg. Calcular la probabilidad de que:

    a) el peso de una caja tomada al azar esté comprendido entre 19.7 y 20.6 kg;

    b) la media de una muestra de 100 cajas esté comprendida entre 19.7 y 20.6 kg.

18. Suponga que $X_1, X_2, \ldots, X_{40}$ representa una muestra aleatoria de mediciones independientes de la proporción de impurezas en muestras de una aleación. Se considera que la densidad de probabilidad de cada una de estas variables aleatorias es $f(x) = 3x^2 \;\; 0 < x < 1$ y $f(x) = 0$ si $x < 0$ o $1 < x$. Un comprador potencial rechaza una partida de mineral si la media de una muestra de tamaño 40 es superior a 0.8. Calcular la probabilidad de que la partida sea rechazada.

19. Se toman muestras independientes de tamaño 10 y 15 de una variable aleatoria con distribución normal de media 20 y varianza 3. ¿Cuál es la probabilidad de que el promedio de las dos muestras se diferencie (en valor absoluto) en más de 0.3?

    Resolución aquí. Video con el ejercicio explicado.

20. Se considera la variable aleatoria X: número obtenido al arrojar un dado no cargado.

    a) Obtener la distribución de probabilidades de X, calcular su valor esperado y su varianza.

    b) Enumerar todas las posibles muestras de tamaño 2 que se pueden obtener de la variable X (son 36 resultados posibles). Calcular para cada una de ellas el promedio y determinar entonces la distribución de probabilidades de ese promedio (media muestral). Compare los valores medios de X y de la media muestra ¿qué concluye?, ¿qué resulta al comparar las varianzas?

    c) Analice la distribución en el muestreo de la varianza muestral S^2. Considere todas las muestras posibles de tamaño 2 y calcule para cada una la varianza muestral. Encuentre la distribución de probabilidades de la varianza muestral y compare su valor esperado con la varianza poblacional.

21. Genere 50 muestras cada una de tamaño 30 de una variable aleatoria con distribución exponencial de parámetro característico 2. Para cada muestra calcular la media muestral obteniéndose así 50 realizaciones del estadístico media muestral para muestras de tamaño 10. Agrupe esos datos y analice las frecuencias relativas. Repita el procedimiento pero ahora con un tamaño de muestra igual a 500. Comente los resultados.

```
R = rand(30,500);E = -(1/2)*log(1-R);
Xraya = mean(E);
Z =(Xraya-0.5)/(0.5/sqrt(30));
ZZ = sort(Z);
plot(ZZ,F,ZZ,normcdf(ZZ))
D = max(abs(F-normcdf(ZZ)));
```

    En una ejecución de este código la distancia `D` fue `0.065007`. Aqui se compara la distribución empírica de los valores generados de `Z` con la distribución normal estándar.

22. En la fabricación de cierto tipo de cojinetes para motor se sabe que el diámetro promedio es de 5 cm con un desvío estándar de 0.005 cm. El proceso es controlado en forma periódica mediante la selección aleatoria de 64 cojinetes, midiendo sus correspondientes diámetros. El proceso se supone bajo control si la media muestral se encuentra entre dos límites especificados en el 95 % de las veces que se extraen las muestras para realizar el control del proceso. Determinar los valores de esos límites si son simétricos respecto de la media poblacional.

23. Los montos X de las facturas tienen una función densidad de probabilidad dada por $\mathrm{f}_X(x) = 3/x^4$ para $x > 1$ y 0 para $x < 1$ (X en centenas de \$). Suponiendo que los montos de facturas diferentes son independientes ¿cuál es la probabilidad de que el monto de 300 facturas del mes supere \$ 42000?

24. Para facilitar el cobro de facturas, una empresa descuenta redondeando a la decena a favor del cliente. La empresa completa 1000 facturas por mes.

    a) ¿Cuál es el monto medio de descuento en el mes por esta causa?

    b) ¿Entre qué valores puede Ud. informar que estará ese valor, con probabilidad 0.95? Considere que el intervalo sea simétrico respecto del valor medio.

    c) ¿Cuál es el máximo descuento mensual tal que pueda asegurarse que no va a ser superado con probabilidad de 0.95?

ACTIVIDAD OPTATIVA (recomendable para fijar conceptos).
Actividades en Octave.
Una verificación del Teorema Central del Límite (TCL)

Dado que toda variable aleatoria con distribución binomial de parámetros n y p puede representarse como suma de n variables aleatorias independientes con distribución Bernoulli de parámetro p, es razonable la idea de aproximar dicha distribución utilizando la de una normal de media np y desviación $\sqrt{np(1-p)}$. Usando algún programa (por ejemplo, Excel, Octave, Matlab o R) represente las funciones de distribución binomial y la de su aproximación por la normal para distintos valores de n y p (por ejemplo $n = 10, 30, 60, 100, 400$ y $p = 0.01, 0.1, 0.25, 0.5, 0.75, 0.9, 0.99$ y obtenga, en forma aproximada, la máxima diferencia en valor absoluto entre la distribución calculada en forma exacta y en forma aproximada en el conjunto $0, 1, \ldots n$. Por ejemplo si $n = 400$ y p =0.5 esa máxima diferencia es 0.0199 (usando Octave).

El siguiente código en *Octave* muestra como obtener 4 gráficos de la distribución binomial para los parámetros (n, p) de estos pares: (5, 0.01), (5, 0.5), (50, 0.01) y (50, 0.5).

```
P1 = binopdf(0:5,5,0.01);P2=binopdf(0:5,5,0.5);
P3 = binopdf(0:50,50,0.01);P4=binopdf(0:50,50,0.5);
subplot(2,2,1),stem(0:5,P1),subplot(2,2,2),stem(0:5,P2)
subplot(2,2,3),stem(0:50,P3),subplot(2,2,4),stem(0:50,P4)
```

El comando `binopdf(x,n,p)` calcula los valores de la función de probabilidad de la variable aleatoria con distribución binomial para los elementos del vector `x` si los parámetros de la distribución son `n` y `p` (el número de repeticiones del experimento de Bernoulli y la probabilidad de éxito respectivamente). El comando de gráficos `stem` se usa para la representación de funciones de variable discreta.

Otra verificación del Teorema Central TCL

Una variable aleatoria con distribución de Poisson de parámetro λ siempre puede representarse como suma de n variables aleatorias Poisson independientes de parámetro λ/n. Por lo tanto es razonable que dicha distribución pueda ser aproximada por la de una normal de media λ y desviación $\sqrt{\lambda}$, sobre todo para grandes valores de λ. Usando algún programa (por ejemplo, Excel, Octave, Matlab o R) represente las funciones de distribución de Poisson y la de su aproximación por la normal para $\lambda = 1, 5, 10, 20, 50, 100$ y obtenga, en forma aproximada, la máxima diferencia en valor absoluto entre la función de distribución calculada en forma exacta y en forma aproximada en el conjunto $0, 1, \ldots 5\,\lambda$. Por ejemplo si $\lambda = 50$ esa máxima distancia es 0.03752 (calculado con Octave).

El siguiente código en *Octave* muestra como obtener 4 gráficos de la distribución de Poisson para el parámetro tomando los valores 1, 3, 5 y 10.

```
P1=poisspdf(0:10,1);P2=poisspdf(0:20,3);
P3=poisspdf(0:20,5);P4=poisspdf(0:40,10);
subplot(2,2,1),stem(0:10,P1),subplot(2,2,2),stem(0:20,P2)
subplot(2,2,3),stem(0:20,P3),subplot(2,2,4),stem(0:40,P4)
```

Se muestra a continuación una aproximación de la distribución de Poisson por la normal y la máxima distancia entre las funciones de distribución si el parámetro de la distribución de Poisson vale 50 y se analiza el intervalo el conjunto $0, 1, \cdots 250$.

```
L = 50;
x = 0:5*L;
poisson = poisscdf(x,L);
E = L;S = sqrt(L);
normal = normcdf((x-E)/S);
max(abs(poisson-normal))
ans = 0.037517
```

El comando `poisscdf(x,L)` calcula los valores de la función de distribución de la variable aleatoria con distibución de Poisson para los elementos del vector `x` si el parámetro de la distribución es `L`.

Otra verificación del Teorema Central del Límite por simulación

Sea $X = X_1 + X_2 + X_3 + X_4 + X_5$, con X_i variables aleatorias independientes con distribución uniforme en $(0, 1)$. Se desea saber si la distribución de X puede o no aproximarse por una normal. Para esto, se propone generar 200 copias simuladas de la variable aleatoria X. Se procede de la siguiente manera usando una planilla de cálculo (Excel, por ejemplo):

1. utilizando la función aleatorio() (o rand()) generar una matriz de 200 x 5 números.
2. Sumar por columnas, generando una 6ta columna, cuyos elementos serán copias simuladas de X.
3. Graficar estos valores en un histograma mediante la opción: Herramientas/Análisis de Datos/Histograma (en Excel)
4. Evaluar si dicho histograma posee o no forma de campana, y en consecuencia responder si la aproximación propuesta es adecuada o no.

En *Octave* se puede ensayar la ejecución de estas líneas:

```
N = 1000; Xs = sum(rand(5,N)); hist(Xs)
XX=sort(Xs);F=(1:N)/N;
mu=2.5; s=sqrt(5/12); FN=normcdf(XX,mu,s);
plot(XX,FN,XX,F)
max(abs(FN-F))
```

Con este código se genera el histograma para `N` valores simulados de la suma de 5 variables aleatorias uniformes independientes. La instrucción `rand(5,N)` genera una matriz $5 \times N$ con números pseudoaleatorios de distribución uniforme en $(0, 1)$, mientras que `Xs` almacena la suma de cada columna de la matriz. De ese modo se obtienen N valores simulados de X, cuyo histograma es dado por la instrucción `hist(Xs)`.

El vector `XX` contiene las componentes de `Xs` ordenadas de menor a mayor. Si a cada valor del vector `XX` se le asigna la probabilidad 1/N se obtiene una variable aleatoria discreta, cuya función de distribución se denomina distribución empírica de la variable aleatoria X y que cuando N es grande, aproxima bien a la distribución de X. El vector `F` contiene los valores de esa distribución empírica en cada uno de los valores almacenados en el vector `XX`, en el mismo orden. Por otra parte `FN` es un vector que almacena los valores de la función de distribución normal de media $\mu = \mathrm{E}[X]$ y desviación $\sigma = \sqrt{\mathrm{Var}[X]}$ evaluada en las componentes de `XX` y en el mismo orden (aproximación de X mediante una normal con mismas media y desviación). Finalmente `plot(XX,FN,XX,F)` produce los gráficos de `F` y `FN` en una misma figura, y `max(abs(FN-F))` calcula el máximo apartamiento entre esos dos gráficos. Si desea, por ejemplo el gráfico de `F`, ejecute `plot(XX,F)`. Ejecutando este código podrá ver que los gráficos aparecen superpuestos a la vista. Para una ejecución esa distancia resultó 0.019441.

ACTIVIDAD OPTATIVA (recomendable para fijar conceptos).
Sobre distribuciones en el muestreo, usando *Statlets*, el laboratorio virtual de la Universidad de Rice:

`http://onlinestatbook.com/stat_sim/sampling_dist/index.html`

Haga click sobre el botón `Begin` y el applet de Java se ejecutará en una nueva ventana. Este applet de Java explora varios aspectos de la distribución en el muestreo de un estadístico o estimador. Cuando se ejecuta por primera vez puede verse un histograma de la distribución normal que es la distribución de probabilidades de la variable aleatoria definida en la población de la que se extrae la muestra. La media de la distribución de probabilidades de la variable definida en la población se indica con una pequeña linea azul y la mediana con una roja, como para la distribución normal son iguales las dos marcas están superpuestas. La línea roja que se extiende desde la media hacia cada lado corresponde al intervalo de semiamplitud igual al desvío estándar (poblacional en este caso)con centro en la media. El segundo histograma muestra los datos muestrales, inicialmente está en blanco. El tercer y cuarto histogramas muestran la distribución de probabilidades de los estadísticos obtenidos a partir de las muestra. El número de veces (replicaciones) que se extraen muestras de tamaño N se indica con el valor de la variable Reps=."

Operaciones básicas

Inicialmente la simulación corresponde a la extracción de una muestra de tamaño N = 5, se calcula la media y se representa su valor. Activando el botón `Animated sample` se pueden ver las cinco observaciones muestrales. Se calcula la media de esos cinco números y se representa ese valor en el tercer histograma. Es interesante repetir esta operación varias veces y observar como se va obteniendo una distribución de valores de la media muestral.Puede acelerarse el proceso tomando 5, 100, 1000, o 10000 muestras.

Elección del estadístico

Los siguientes estadísticos pueden ser calculados para cada muestra extraída: Media, Desvío estándar de la muestra (usando N en el denominador) Varianza de la muestra (usando N en el denominador) Estimador insesgado de la varianza (usando N-1 en el denominador) Valor absoluto del desvío respecto de la media. Rango.

Elección del tamaño muestra

El tamaño de la muestra puede elegirse en 2, 5, 10, 16, 20 or 25 en el menú. Observe no confundir el tamaño de la muestra (N) con el número de muestras de ese tamaño que son generadas para poder observar la distribución en el muestreo del estadístico elegido.

Comparación con la distribución normal

Pulsando el botón `Fit normal` puede verse la superposición del gráfico de la función densidad de la distribución normal sobre el histograma de la distribución en el muestreo analizada.

5. Respuestas

1. Por la desigualdad de Markov, la probabilidad $\leq 5/6 = 0.8333$.
2.

1. 0.2515.
2. 0.2942.
3. 0.0228.
4. 0.0003.

3. Con la distribución binomial 0.9831, con la aproximación normal 0.9825.

4. Con la distribución binomial 0.9601, con la aproximación normal 0.9664.

5. 78.

6. a) 0.151 b) 0.309 c) 0.039 (3.9 %).

7. 0.4276.

8. 0.1802.

9. a) 0.259 b) 0.251.

10. Con la distribución de Poisson (con Octave) da 0.1719. Con la aproximación normal y con la corrección por continuidad (y usando Octave) da 0.1725.

11. a) 0.76 b) 0.59 c) 3002.3 litros.

12. a) 0.00135 b) 88.

14. a) $n = 3523$, tomando $\Phi^{-1}(0.99) = 2.3263$ b) $n \leq 13529$.

15. a) Mayor que 0.8200 (usando la desigualdad de Tchebycheff). Usando la distribución binomial la probabilidad es 0.9703.
b) Menor que 0.1250 (usando la desigualdad de Tchebycheff). Usando la distribución binomial la probabilidad es 0.0094.

16. Usando la desigualdad de Tchebycheff da 3333. Asumiendo que la frecuencia relativa tiene distribución normal la respuesta es 471.

17. a) 0.1191 b) 0.8185.

18. 0.051.

19. 0.6714.

22. (4.999 , 5.001).

23. 0.9773.

24. Si se supone que el descuento por factura se supone una variable aleatoria discreta uniforme con recorrido $\{1, 2, 3, 4, 5, 6, 7, 8, 9\}$:

a) El monto de descuento en un mes es una variable aleatoria con distribución aproximadamente normal (en el sentido del TCL) de media 5000 y desvío estándar 81.65.
b) Si ese intervalo I está centrado en la media entonces $I = (4840, 5160)$.
c) \$5134 .

Si se supone que el descuento por factura se supone una variable aleatoria continua uniforme (0, 10):
a) El monto de descuento en un mes es una variable aleatoria con distribución aproximadamente normal (en el sentido del TCL) de media 5000 y desvío estándar 91.29.
b) Si ese intervalo I está centrado en la media entonces $I = (4821, 5479)$.
c) \$5150.

6. Ejercicios resueltos

Ejercicio 1

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.

Resolución

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

$$\mathrm{P}(70 < X < 90) = \mathrm{P}(71 \leq X \leq 89) = \sum_{k=71}^{89} \binom{100}{k} 0.8^k 0.2^{100-k} \approx 0.98305. \tag{1}$$

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:

$$\mathrm{P}(70 < X < 90) = \mathrm{P}(70.5 < X \leq 89.5) \approx \Phi\left(\frac{89.5 - 80}{4}\right) - \Phi\left(\frac{70.5 - 80}{4}\right) \approx 0.98245. \tag{2}$$

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,

$$\mathrm{P}(70 < X < 90) = \mathrm{P}(|X - 80| \leq 9) = 1 - \mathrm{P}(|X - \mathrm{E}[X]| \geq 10) \geq 1 - \frac{16}{10^2} = 0.84. \tag{3}$$

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

Ejercicio 3

Una central telefónica A da servicio a 1800 usuarios de una central cercana B. Sería costoso y extravagante instalar 1800 líneas troncales de A a B. Es suficiente instalar un número N de líneas tan grande que, en condiciones ordinarias, solamente una de cada 100 llamadas en promedio no encuentre inmediatamente una línea troncal disponible. Supóngase que, en la hora más ocupada del día, cada usuario requiere una linea troncal de B durante un promedio de 2 minutos. En un momento fijo de la hora de máximo tráfico, puede suponerse la situación como un conjunto de 1800 ensayos independientes con una probabilidad $p = \frac{1}{30}$ (por lo de 2 min de cada 60 min) en que cada uno se requiere una línea. Determinar el número de líneas N a instalar entre A y B. Explicar claramente el planteo.

Resolución

Variables aleatorias

En este ejercicio, podemos pensar en la siguiente variable aleatoria, pensando en una línea por usuario:

$$S_{1800} = \text{cantidad de usuarios que requieren una línea en la hora más ocupada del día} \tag{4}$$

Datos

Particularmente, podríamos pensar a la variable S_{1800} como una binomial de parámetros 1800 y $p = \frac{1}{30}$. Sin embargo, para usar el Teorema central del límite, se puede pensar a la variable S_{1800} como una suma de variables independientes Bernoulli:

$$S_{1800} = \sum_{i=1}^{1800} X_i, \quad \text{donde } X_i = \begin{cases} 1 & \text{si el } i\text{-ésimo cliente requiere una línea} \\ 0 & \text{en caso contrario} \end{cases} \tag{5}$$

donde $\mathrm{E}[X_i] = \frac{1}{30}$ y $\sigma(X_i) = \sqrt{\frac{1}{30} \cdot \frac{29}{30}} = \frac{\sqrt{29}}{30}$.

Por lo tanto, al ser una suma de variables aleatorias independientes idénticamente distribuidas, y al ser 1800 suficientemente grande, el teorema central del límite nos dice que S_{1800} se puede aproximar por una variable con distribución normal con los siguientes parámetros:

$$S_{1800} \overset{(a)}{\sim} \mathcal{N}\left(1800 \cdot \mathrm{E}[X_i]; \sqrt{1800} \cdot \sigma(X_i)\right) = \mathcal{N}\left(\frac{1800}{30}; \frac{\sqrt{29 \cdot 1800}}{30}\right) = \mathcal{N}\left(60; \sqrt{58}\right) \Rightarrow \frac{S_{1800} - 60}{\sqrt{58}} \overset{(a)}{\sim} \mathcal{N}(0;1) \tag{6}$$

Cálculo de N

Se pide que en promedio uno de cada 100 llamados no encuentre una línea disponible. Esto se puede pensar como pedir que la probabilidad de sobrepasar las líneas disponibles sea $\frac{1}{100}$, es decir:

$$\mathrm{P}(S_{1800} > N) = \frac{1}{100} \tag{7}$$

Sin embargo, como N debe ser entero, puede no cumplirse esta igualdad y por lo tanto, desde la perspectiva de la compañía el objetivo es **acotar superiormente** la probabilidad de que la demanda sobrepase las líneas disponibles, es decir:

$$\mathrm{P}(S_{1800} > N) \leq \frac{1}{100} \tag{8}$$

Seguiremos asumiendo que la cantidad de líneas es suficientemente grande para poder usar TCL, y en caso de que el resultado final no sea consistente con esta hipótesis, usaremos la distribución binomial. Por lo tanto, debemos usar la corrección por continuidad:

$$\mathrm{P}(S_{1800} > N) = \mathrm{P}(S_{1800} > N + 0.5) \tag{9}$$

Vale aclarar que esa es una igualdad ya que el recorrido de la variable S_N está incluido en los números enteros, y que no se incluye un valor de probabilidad positiva al añadir 0.5. Ahora sí, podemos hacer la aproximación:

$$\mathrm{P}(S_{1800} > N + 0.5) = \mathrm{P}\left(\frac{S_{1800} - 60}{\sqrt{58}} > \frac{N + 0.5 - 60}{\sqrt{58}}\right) \overset{\text{TCL}}{\approx} 1 - \Phi\left(\frac{N - 59.5}{\sqrt{58}}\right) \tag{10}$$

Por lo tanto, podemos buscar N de forma que:

$$1 - \Phi\left(\frac{N - 59.5}{\sqrt{58}}\right) \leq 0.01 \Rightarrow 0.99 \leq \Phi\left(\frac{N - 59.5}{\sqrt{58}}\right) \Rightarrow z_{0.99} \leq \frac{N - 59.5}{\sqrt{58}} \tag{11}$$

Despejando,

$$z_{0.99} \cdot \sqrt{58} + 59.5 \leq N \Rightarrow 77.217 \leq N \tag{12}$$

Por lo tanto, como N debe ser entero, tomamos $N = 78$. Notar que es sensiblemente menor al valor inicial 1800, de forma que se reducen considerablemente los costos asociados a la instalación de líneas.

Verificación

Hoy en día, podemos usar software para verificar estas cuentas de forma exacta, utilizando la distribución binomial de S_{1800}. Por ejemplo, en este caso, podríamos verificar que:

$$\mathrm{P}(S_{1800} > 78) = 0.0096038 \leq 0.01 \tag{13}$$

Por lo tanto, se cumple lo pedido, pero más aún, vemos que **no** se cumple lo pedido para $N = 77$

$$\mathrm{P}(S_{1800} > 77) = 0.013178 \not\leq 0.01 \tag{14}$$

por lo que podemos inferir que la aproximación es suficientemente precisa.

Ejercicio 5

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.

Resolución

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

$$\mathrm{P}(T > 310) = 1 - \mathrm{P}(T \leq 310). \tag{15}$$

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

$$\mathrm{P}(T > 310) = \mathrm{P}(N < 30) = \sum_{k=0}^{29} \frac{31^k}{k!} e^{-31}. \tag{16}$$

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).

Resolución usando las desigualdades de Markov y Chebychev

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

$$\mathrm{P}(T > 310) \leq \frac{300}{310} = \frac{30}{31} \approx 0.9677. \tag{17}$$

La cota es válida, pero no es muy ajustada (ver resultado más arriba).

También sabemos que $\mathrm{Var}[T] = 30 \times 1/0.1^2 = 3000$. Utilizando la desigualdad de Chebychev,

$$\mathrm{P}(T > 310) = \mathrm{P}(T - 300 > 10) \leq \mathrm{P}(|T - 300| > 10) \leq \frac{3000}{10^2} = 30 > 1. \tag{18}$$

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.

Resolución usando el Teorema Central del Límite

Sabemos que $\mathrm{E}[T] = 30 \times 1/0.1 = 300$ 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

$$\mathrm{P}(T > 310) = 1 - \mathrm{P}(T \leq 310) \approx 1 - \Phi\left(\frac{310 - 300}{\sqrt{3000}}\right) = 1 - \Phi\left(\frac{1}{\sqrt{30}}\right) \approx 0.4276. \tag{19}$$

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

Ejercicio 12

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.

Resolución

El ejercicio pide encontrar n tal que

$$\mathrm{P}\left(\left|\hat{P}_n - p\right| < 0.01\right) \geq 0.98. \tag{20}$$

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 $\mu_n = \mathrm{E}[X_n] = np$ y $\sigma_n^2 = \mathrm{Var}[X_n] = np(1-p)$. Luego, el ejercicio pide encontrar n tal que

$$\mathrm{P}\left(\left|\hat{P}_n - p\right| < 0.01\right) = \mathrm{P}\left(|X_n - np| < 0.01n\right) = \sum_{k=\lceil \mu_n - 0.01n \rceil}^{\lfloor \mu_n + 0.01n \rfloor} \binom{n}{k} p^k (1-p)^{n-k} \geq 0.98. \tag{21}$$

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,

$$\begin{aligned} 0.98 \leq& \mathrm{P}\left(\left|\hat{P}_n - p\right| < 0.01\right) \approx \\ &\Phi\left(\frac{\lfloor \mu_n + 0.01n \rfloor + 0.5 - \mu_n}{\sigma_n}\right) - \Phi\left(\frac{\lceil \mu_n - 0.01n \rceil - 0.5 - \mu_n}{\sigma_n}\right), \end{aligned} \tag{22}$$

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

$$0.98 \leq \mathrm{P}\left(\left|\hat{P}_n - p\right| < 0.01\right) \approx \Phi\left(\frac{0.01n}{\sigma_n}\right) - \Phi\left(\frac{-0.01n}{\sigma_n}\right) = 2\Phi\left(\frac{0.01n}{\sigma_n}\right) - 1. \tag{23}$$

Basados en la aproximación, buscaremos n tal que

$$\Phi\left(\frac{0.01n}{\sigma_n}\right) \geq \frac{1 + 0.98}{2} = 0.99. \tag{24}$$

Luego,

$$\frac{0.01n}{\sigma_n} \geq z_{0.99} \Rightarrow 0.01n \geq z_{0.99}\sigma_n = z_{0.99}\sqrt{np(1-p)} \Rightarrow \sqrt{n} \geq \frac{z_{0.99}\sqrt{p(1-p)}}{0.01} \Rightarrow \tag{25}$$

$$n \geq 10000\, z_{0.99}^2 \left[p(1-p)\right]. \tag{26}$$

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

$$n \geq 3523.1 \Rightarrow n \geq 3524. \tag{27}$$

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

$$\mathrm{P}\left(\left|\hat{P}_{3524} - 0.07\right| < 0.01\right) = \mathrm{P}\left(211 < X_{3524} \leq 281\right) \approx 0.9792 \approx 0.98. \tag{28}$$

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.

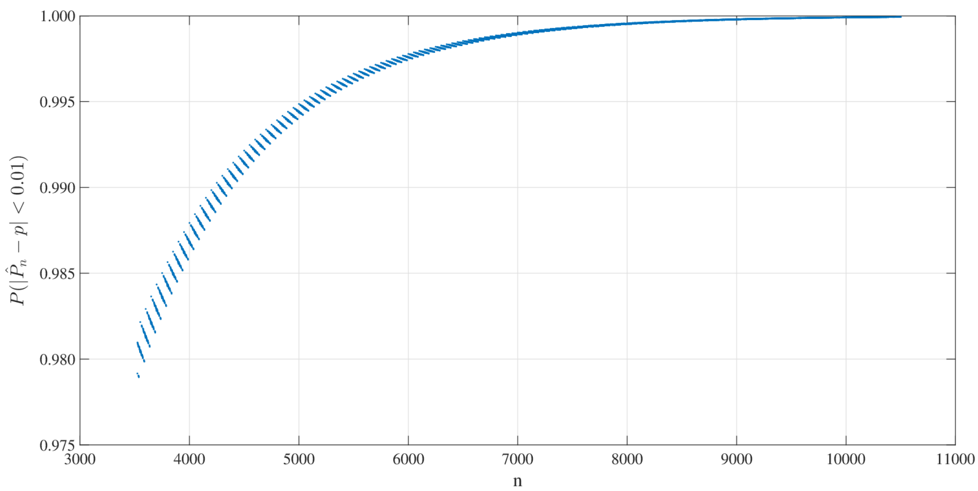


Figura 1: Respuesta al ejercicio 12 de la guía 7 cuando $p = 0.07$.

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

$$n \geq 10000\, z_{0.99}^2 \left[p(1-p)\right]. \tag{29}$$

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

$$n \geq 10000\, z_{0.99}^2 \max_{p\in[0,1]} \left[p(1-p)\right]. \tag{30}$$

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

$$g'(p) = 1 - 2p, \qquad g''(p) = -2. \tag{31}$$

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

$$n \geq 10000\, z_{0.99}^2 \, \frac{1}{4} = 13529.74 \Rightarrow n \geq 13530. \tag{32}$$

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.

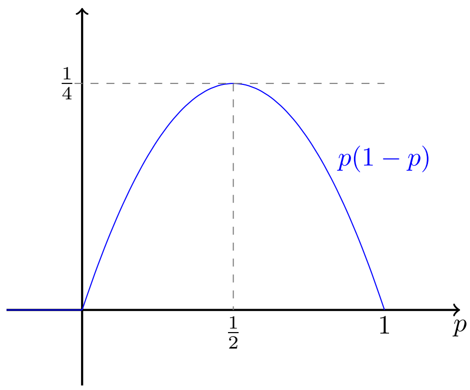


Figura 2: Varianza de una variable aleatoria Bernoulli: $p(1-p)$.

Ejercicio 14

Se desea estimar la probabilidad p de ocurrencia de cierto suceso a partir de la frecuencia relativa de ocurrencia medida al realizar n repeticiones del experimento en que puede ocurrir el suceso de interés. Se pone como condición que la probabilidad de que el error (diferencia entre p y la frecuencia relativa) no supere el valor 0.05, sea mayor que 0.97. ¿Cuántas pruebas deben realizarse como mínimo? Comparar los resultados obtenidos a partir de usar:

a) la desigualdad de Tchebycheff.

b) la aproximación normal de la distribución binomial.

Resolución

Variables aleatorias

El enunciado hace mención a la frecuencia relativa de ocurrencia del suceso de interés. La pregunta es cómo representarlo en términos de una variable aleatoria. Podríamos definir la frecuencia **absoluta** de ocurrencia del suceso del siguiente modo:

$$X_n = \text{ cantidad de veces que ocurre el suceso de interés en } n \text{ repeticiones del experimento} \tag{33}$$

Por lo tanto, la frecuencia **relativa** vendrá dada por el cociente entre la cantidad absoluta de ocurrencias (X_n) y la cantidad de repeticiones (n):

$$\mathrm{f}_R = \frac{X_n}{n} \tag{34}$$

Datos

Al ser independientes las n repeticiones del experimento con la misma probabilidad p de éxito, $X_n \sim \mathrm{Bi}(n, p)$. Por lo tanto, podemos inferir que $\mathrm{E}[X_n] = n \cdot p$ y $\mathrm{Var}[X_n] = n \cdot p \cdot (1-p)$. A partir de estos datos, podemos calcular el valor esperado y varianza de la frecuencia relativa:

- $\mathrm{E}[\mathrm{f}_R] = \mathrm{E}\left[\frac{X_n}{n}\right] = \frac{\mathrm{E}[X_n]}{n} = \frac{\cancel{n} \cdot p}{\cancel{n}} = p$
- $\mathrm{Var}[\mathrm{f}_R] = V\left(\frac{X_n}{n}\right) = \frac{\mathrm{Var}[X_n]}{n^2} = \frac{\cancel{n} \cdot p \cdot (1-p)}{n^{\cancel{2}}} = \frac{p \cdot (1-p)}{n}$

Ítem a

Recordemos la desigualdad de Tchebychev:

$$\mathrm{P}(|X - \mathrm{E}[X]| \geq \varepsilon) \leq \frac{\mathrm{Var}[X]}{\varepsilon^2} \tag{35}$$

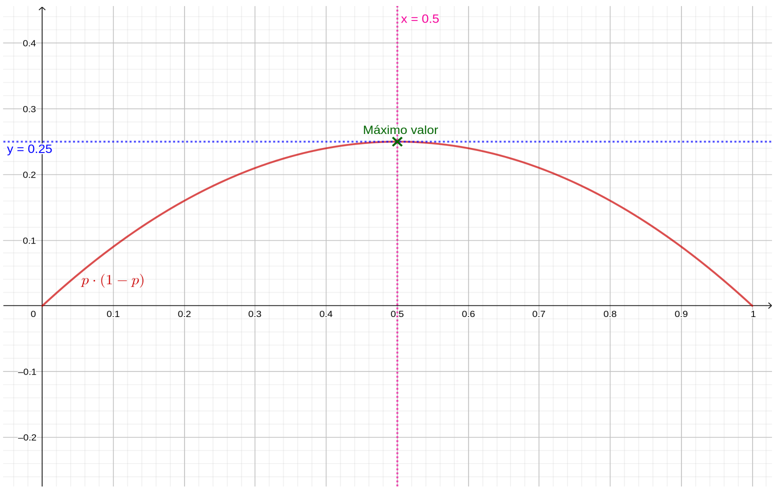
En este caso, nos piden n de forma que:

$$\mathrm{P}(|\mathrm{f}_R - p| \leq 0.05) > 0.97 \Leftrightarrow 1 - \mathrm{P}(|\mathrm{f}_R - p| > 0.05) > 0.97 \Leftrightarrow \mathrm{P}(|\mathrm{f}_R - p| > 0.05) < 0.03 \tag{36}$$

Como $\mathrm{E}[\mathrm{f}_R] = p$, podemos aplicar la desigualdad de Tchebychev con $\varepsilon = 0.05$. Por otro lado, a medida que n se hace grande, la probabilidad de que $|\mathrm{f}_R - p| = 0.05$ se achica (más aún, es cero si $n \cdot (p + 0.05)$ y $n \cdot (p - 0.05)$ no son enteros) y por lo tanto, podemos considerarlo despreciable:

$$\mathrm{P}(|\mathrm{f}_R - p| > 0.05) = \mathrm{P}(|\mathrm{f}_R - Ep| \geq 0.05) = \mathrm{P}(|\mathrm{f}_R - \mathrm{E}[\mathrm{f}_R]| \geq 0.05) = \leq \frac{\mathrm{Var}[\mathrm{f}]_R)}{0.05^2} = \frac{p \cdot (1-p)}{n \cdot 0.05^2} \tag{37}$$

Si nos aseguramos que este valor esté por debajo de 0.03, se cumple lo pedido. El problema es que no tenemos un valor para p. Por lo tanto, usaremos una cota superior de la expresión correspondiente a p para sólo lidiar con el tamaño muestral n. Como $p \in (0,1)$, $p(1-p)$ tiene un valor máximo de $\frac{1}{4}$ (cuando $p = \frac{1}{2}$):



Por lo tanto,

$$P(|f_R - p| > 0.05) \leq \frac{p \cdot (1-p)}{n \cdot 0.05^2} \leq \frac{\frac{1}{4}}{n \cdot 0.05^2} = \frac{1}{4n \cdot 0.05^2} \tag{38}$$

Es decir, si

$$\frac{1}{4n \cdot 0.05^2} < 0.03 \Rightarrow P(|f_R - p| > 0.05) \leq \frac{1}{4n \cdot 0.05^2} < 0.03 \Rightarrow P(|f_R - p| > 0.05) < 0.03 \tag{39}$$

Por lo tanto, buscaremos n a partir de la primer inecuación, ya que en ese caso, se cumple lo pedido:

$$\frac{1}{4n \cdot 0.05^2} < 0.03 \Rightarrow \frac{1}{4 \cdot 0.03 \cdot 0.05^2} < n \Rightarrow n > 3333.3 \Rightarrow n \geq 3334 \tag{40}$$

Ítem b

Como X_n es binomial, si n es suficientemente grande, a través del TCL podemos aproximar su distribución por una normal:

$$X_n \overset{(a)}{\sim} \mathcal{N}\left(n \cdot p; \sqrt{n \cdot p \cdot (1-p)}\right) \Rightarrow f_R = \frac{X_n}{n} \overset{(a)}{\sim} \mathcal{N}\left(p; \sqrt{\frac{p \cdot (1-p)}{n}}\right) \Rightarrow \frac{f_R - p}{\sqrt{\frac{p \cdot (1-p)}{n}}} \overset{(a)}{\sim} \mathcal{N}(0,1) \tag{41}$$

Por lo tanto,

$$\begin{aligned}
\mathrm{P}(|\mathrm{f}_R - p| \leq 0.05) = \mathrm{P}(-0.05 \leq \mathrm{f}_R - p \leq 0.05) &= \mathrm{P}\left(-\frac{0.05}{\sqrt{\frac{p\cdot(1-p)}{n}}} \leq \frac{\mathrm{f}_R - p}{\sqrt{\frac{p\cdot(1-p)}{n}}} \leq \frac{0.05}{\sqrt{\frac{p\cdot(1-p)}{n}}}\right) \\
&\approx \Phi\left(\frac{0.05}{\sqrt{\frac{p\cdot(1-p)}{n}}}\right) - \Phi\left(-\frac{0.05}{\sqrt{\frac{p\cdot(1-p)}{n}}}\right) \\
&\approx \Phi\left(\frac{0.05\sqrt{n}}{\sqrt{p\cdot(1-p)}}\right) - \left(1 - \Phi\left(\frac{0.05\sqrt{n}}{\sqrt{p\cdot(1-p)}}\right)\right) \\
&\approx 2\cdot\Phi\left(\frac{0.05\sqrt{n}}{\sqrt{p\cdot(1-p)}}\right) - 1 \qquad (42)
\end{aligned}$$

Luego, para que se cumpla lo pedido, podemos usar la aproximación:

$$\begin{aligned}
\mathrm{P}(|\mathrm{f}_R - p| \leq 0.05) \approx & 2\cdot\Phi\left(\frac{0.05\sqrt{n}}{\sqrt{p\cdot(1-p)}}\right) - 1 > 0.97 \Leftrightarrow \\
\Leftrightarrow & \Phi\left(\frac{0.05\sqrt{n}}{\sqrt{p\cdot(1-p)}}\right) > \frac{0.97+1}{2} = 0.985 \qquad (43)
\end{aligned}$$

$$\Leftrightarrow \frac{0.05\sqrt{n}}{\sqrt{p\cdot(1-p)}} > z_{0.985} \Leftrightarrow \sqrt{n} > \frac{z_{0.985}\cdot\sqrt{p\cdot(1-p)}}{0.05} \qquad (44)$$

Nuevamente, como $p\cdot(1-p) \leq \frac{1}{4}$, si se cumple:

$$\sqrt{n} > \frac{z_{0.985}\cdot\sqrt{\frac{1}{4}}}{0.05} \geq \frac{z_{0.985}\cdot\sqrt{p\cdot(1-p)}}{0.05} \Rightarrow \sqrt{n} > \frac{z_{0.985}\cdot\sqrt{p\cdot(1-p)}}{0.05} \qquad (45)$$

Por lo tanto, buscamos n de forma que:

$$\sqrt{n} > \frac{z_{0.985}\cdot\sqrt{\frac{1}{4}}}{0.05} \Rightarrow n > \left(\frac{z_{0.985}}{2\cdot 0.05}\right)^2 \approx 470.93 \Rightarrow n \geq 471 \qquad (46)$$

Ejercicio 17

Se toman muestras independientes de tamaño 10 y 15 de una variable aleatoria con distribución normal de media 20 y varianza 3. ¿Cuál es la probabilidad de que el promedio de las dos muestras se diferencie (en valor absoluto) en más de 0.3?

Resolución

Variables aleatorias

Las muestras independientes provienen de variables idénticamente distribuidas (podemos llamarlas D_i, E_i), es decir, podemos considerar:

$$D_i, E_i \sim \mathcal{N}(20, \sqrt{3}), \quad \overline{X}_{10} = \sum_{i=1}^{10} \frac{D_i}{10}, \quad \overline{Y}_{15} = \sum_{i=1}^{15} \frac{E_i}{15} \tag{47}$$

Datos

Al ser D_i y E_i normales independientes, sus promedios también son normales. Sólo resta determinar sus parámetros:

- $\mathrm{E}\left[\overline{X}_{10}\right] = \mathrm{E}\left[\sum_{i=1}^{10} \frac{D_i}{10}\right] = \sum_{i=1}^{10} \frac{\mathrm{E}\left[D_i\right]}{10} = \sum_{i=1}^{10} \frac{20}{10} = \frac{10 \cdot 20}{10} = 20$
- $\mathrm{Var}\left[\overline{X}_{10}\right] = \mathrm{Var}\left[\sum_{i=1}^{10} \frac{D_i}{10}\right] \overset{IND}{=} \sum_{i=1}^{10} \mathrm{Var}\left[\frac{D_i}{10}\right] = \sum_{i=1}^{10} \frac{\mathrm{Var}\left[D_i\right]}{10^2} = \sum_{i=1}^{10} \frac{3}{100} = \frac{10 \cdot 3}{100} = \frac{3}{10} \Rightarrow \sigma(\overline{X}_{10}) = \sqrt{0.3}$

Por lo tanto $\overline{X}_{10} \sim \mathcal{N}\left(20, \sqrt{\frac{3}{10}}\right)$. Por otro lado:

- $\mathrm{E}\left[\overline{Y}_{15}\right] = \mathrm{E}\left[\sum_{i=1}^{15} \frac{E_i}{10}\right] = \sum_{i=1}^{15} \frac{\mathrm{E}\left[E_i\right]}{15} = \sum_{i=1}^{15} \frac{20}{15} = \frac{15 \cdot 20}{15} = 20$
- $\mathrm{Var}\left[\overline{Y}_{15}\right] = \mathrm{Var}\left[\sum_{i=1}^{15} \frac{E_i}{15}\right] \overset{IND}{=} \sum_{i=1}^{15} \mathrm{Var}\left[\frac{E_i}{15}\right] = \sum_{i=1}^{15} \frac{\mathrm{Var}\left[E_i\right]}{15^2} = \sum_{i=1}^{15} \frac{3}{225} = \frac{15 \cdot 3}{225} = \frac{1}{5} \Rightarrow \sigma(\overline{Y}_{15}) = \sqrt{0.2}$

Luego, $\overline{Y}_{15} \sim \mathcal{N}\left(20, \sqrt{\frac{1}{5}}\right)$.

Nos piden la siguiente probabilidad:

$$\mathrm{P}\left(\left|\overline{X}_{10} - \overline{Y}_{15}\right| > 0.3\right) = 1 - \mathrm{P}\left(\left|\overline{X}_{10} - \overline{Y}_{15}\right| \leq 0.3\right) = 1 - \mathrm{P}\left(-0.3 \leq \overline{X}_{10} - \overline{Y}_{15} \leq 0.3\right) \tag{48}$$

Por lo tanto, puede ser de utilidad saber la distribución de la resta entre $\overline{X}_{10}$ y $\overline{Y}_{15}$. Nuevamente, al ser independientes y normales, su resta es normal y para terminar de conocer su distribución, necesitamos calcular su media y desvío:

- $\mathrm{E}\left[\overline{X}_{10} - \overline{Y}_{15}\right] = \mathrm{E}\left[\overline{X}_{10}\right] - \mathrm{E}\left[\overline{Y}_{15}\right] = 20 - 20 = 0$
- $\mathrm{Var}\left[\overline{X}_{10} - \overline{Y}_{15}\right] \overset{IND}{=} \mathrm{Var}\left[\overline{X}_{10}\right] + \mathrm{Var}\left[-\overline{Y}_{15}\right] = \mathrm{Var}\left[\overline{X}_{10}\right] + \cancel{(-1)^2} \cdot \mathrm{Var}\left[\overline{Y}_{15}\right] = \frac{3}{10} + \frac{1}{5} = \frac{1}{2} \Rightarrow \sigma(\overline{X}_{10} - \overline{Y}_{15}) = \sqrt{\frac{1}{2}}$

Por lo tanto, la probabilidad pedida es:

$$1 - \mathrm{P}\left(-0.3 \leq \overline{X}_{10} - \overline{Y}_{15} \leq 0.3\right) = 1 - \mathrm{P}\left(\frac{-0.3}{\sqrt{\frac{1}{2}}} \leq \frac{\overline{X}_{10} - \overline{Y}_{15}}{\sqrt{\frac{1}{2}}} \leq \frac{0.3}{\sqrt{\frac{1}{2}}}\right) = \tag{49}$$

$$= 1 - \left(\Phi\left(\frac{0.3}{\sqrt{\frac{1}{2}}}\right) - \Phi\left(-\frac{0.3}{\sqrt{\frac{1}{2}}}\right)\right) \overset{\text{simetría de } \Phi}{=} 1 - \left(\Phi\left(0.3 \cdot \sqrt{2}\right) - \left[1 - \Phi\left(0.3 \cdot \sqrt{2}\right)\right]\right) = \tag{50}$$

$$= 1 - \left(2 \cdot \Phi\left(0.3 \cdot \sqrt{2}\right) - 1\right) = 2 - 2 \cdot \Phi\left(0.3 \cdot \sqrt{2}\right) \approx 0.67137 \tag{51}$$

Otra forma

Otra forma de hacerlo es sin tomar el complemento, es decir:

$$\mathrm{P}\left(\left|\overline{X}_{10} - \overline{Y}_{15}\right| > 0.3\right) = \mathrm{P}(\overline{X}_{10} - \overline{Y}_{15} < -0.3) + \mathrm{P}(\overline{X}_{10} - \overline{Y}_{15} > 0.3) \tag{52}$$

Usando que la diferencia entre $\overline{X}_{10}$ y $\overline{Y}_{15}$ tiene media cero y desvío $\sqrt{\frac{1}{2}}$:

$$\mathrm{P}(\overline{X}_{10} - \overline{Y}_{15} < -0.3) + \mathrm{P}(\overline{X}_{10} - \overline{Y}_{15} > 0.3) = \Phi\left(-\frac{0.3}{\sqrt{\frac{1}{2}}}\right) + 1 - \Phi\left(\frac{0.3}{\sqrt{\frac{1}{2}}}\right) = \tag{53}$$

$$= 1 - \Phi\left(0.3 \cdot \sqrt{2}\right) + 1 - \Phi\left(0.3 \cdot \sqrt{2}\right) = 2 \cdot (1 - \Phi\left(0.3 \cdot \sqrt{2}\right)) \approx 0.67137 \tag{54}$$