Estadística en Física Experimental (1^{er} Cuatrimestre 2025)

Guía de Problemas N° 5 | Teorema Central del Límite y Distribución Multinormal

1. Estudie el grado de validez del teorema central del límite dibujando las distribuciones siguientes, y superponiendo sobre ellas la gaussiana con el μ y σ correspondiente.
   (a) $\mathrm{B}_k(5{,}0.2)$, $\mathrm{B}_k(30{,}0.4)$
   (b) $\mathrm{P}_n(4)$, $\mathrm{P}_n(10)$, $\mathrm{P}_n(40)$

2. El teorema central del límite permite evaluar probabilidades binomiales sin necesidad de sumar muchos términos que involucran factoriales de grandes números, a partir de la distribución acumulativa normal canónica $\Phi(x)$,
$$\sum_{k=a}^{b} B_k(n,p) = \sum_{k=a}^{b} \binom{n}{k} p^k q^{n-k} \simeq \Phi\left(\frac{b-np+\frac{1}{2}}{\sqrt{npq}}\right) - \Phi\left(\frac{a-np-\frac{1}{2}}{\sqrt{npq}}\right)$$
   Discuta el origen de esta fórmula y utilícela para calcular la probabilidad de aprobar un examen multiple choice con 100 preguntas de tres opciones cada una, si se contesta al azar y se aprueba con 4 (40% de respuestas correctas). [Rta: 0.0966 con la suma exacta, y 0.0951 con la fórmula aproximada.]

3. Utilizando el teorema central del límite escribir un generador aproximado de números gaussianos N(0,1), a partir de variables aleatorias independientes $\{X_i\}$ con distribución uniforme en [0,1], como una función $f(Z)$ siendo $Z = \sum_i^n X_i$.
   (a) Si se elige n=50, ¿cuál debe ser $f(Z)$?
   (b) ¿En qué rango de la abscisa seguro falla la aproximación a la normal?
   (c) Genere de este modo 10000 números con la computadora, haga un histograma de su distribución, y grafique N(0,1) sobre éste.
   (d) Muestre que el promedio de N variables independientes con distribución de Cauchy tiene a su vez distribución de Cauchy. ¿Por qué falla en este caso el teorema central del límite?

4. ¿Cuánta gente deberá encuestarse en Argentina si se desea conocer la intención de voto p para un cierto candidato dentro de un margen de 1% (en sentido absoluto) y con un nivel de confianza de 95%? Use para esto el dato de que aproximadamente (a) el 45% (b) el 5% del electorado votará efectivamente por dicho candidato. Discuta intuitivamente por qué obtiene resultados distintos para los casos (a) y (b). [Rta: 9900 y 1900]
   Sugerencia: considerar que la población tiene muchos más individuos que cualquiera de estas muestras y usar la aproximación gaussiana.

5. Muestre que a distribución poissoniana tiende a la gaussiana en el límite $\mu \to \infty$. Para ello obtenga la función característica de $Y \equiv (n - E(n))/\sigma_n$, con n poissoniana, y verifique la validez de $\lim_{\mu\to\infty}\phi_Y(t) = \phi_X(t)$, con $X(t)$ gaussiana canónica.

6. Siendo que en el problema anterior no hay una suma de variables aleatorias ¿Por qué esta en esta guía?
   Ayuda: Estudie la distribución de la suma de variables aleatorias con distribución de Poisson.

7. La *Distribución Multinormal* es la generalización a n dimensiones de la normal (la gaussiana) y, al igual que ésta, juega un rol preponderante en probabilidades y estadística. Dadas n variables aleatorias correlacionadas $\{X_i\}$, con esperanza $E(X_i) = \mu_i$ y matriz de covarianza $\mathbb{V}$, ésto es $\mathrm{Cov}(X_i,X_j)=V_{ij}$, se dice que su densidad de probabilidad conjunta $f(\underline{x})$ es multinormal si todas las distribuciones marginales $f(x_i)$ y todas las distribuciones condicionales unidimensionales $f(x_i|x_j, j \neq i)$ son gaussianas. La densidad de probabilidad conjunta $f(\underline{x})$ viene dada por
$$f(\underline{x}) = \frac{1}{\sqrt{(2\pi)^n|\mathbb{V}|}} \exp\left[-\frac{1}{2}\,(\underline{x}-\underline{\mu})^T\mathbb{V}^{-1}(\underline{x}-\underline{\mu})\right]$$
   donde $\underline{x}$ y $\underline{\mu}$ son vectores columna de tamaño n, $\underline{x}^T$ y $\underline{\mu}^T$ los respectivos vectores traspuestos (vectores fila) y $\mathbb{V}$ es cuadrada (de $n{\times}n$), simétrica y definida positiva, con $|\mathbb{V}| \equiv \det(\mathbb{V})$.
   (a) Verifique que para $n = 1$, $f(\underline{x})$ es una gaussiana.

(b) En el caso $n = 2$ (multinormal bivariada) la matriz de covarianza de una multinormal depende de tres parámetros (¿por qué?). Elijamos σ_1, σ_2 y el coeficiente de correlación ρ, ésto es, $V_{11} = \sigma_1^2$, $V_{22} = \sigma_2^2$ y $V_{12} = \rho\,\sigma_1\sigma_2$. Muestre entonces que

$$f(x_1, x_2) = \left(2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}\right)^{-1} \exp\left(-\frac{Q}{2}\right)$$

con

$$Q = \frac{1}{1-\rho^2}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)\right]$$

(c) Compruebe que cuando $\rho = 0$, $f(x_1, x_2) = N(\mu_1, \sigma_1)N(\mu_2, \sigma_2)$. Esto es, para la multinormal, correlación nula implica que las variables son independientes.

En adelante, puede trabajar con $\mu_1 = \mu_2 = 0$ para simplificar las cuentas.

(d) Muestre que la distribución marginal $f(x_2)$ es la gaussiana $N(\mu_2, \sigma_2)$, independientemente del valor del nivel de correlación ρ.

(e) Una manera de visualizar la forma de una multinormal con $n = 2$ es dibujar curvas de nivel de f en el plano x_1, x_2. Considere las correspondientes a $Q = 1$, y muestre que son elipses centradas en (μ_1, μ_2), denominadas *elipses de covarianza.* Para $\mu_1 = \mu_2 = 0$, verifique que éstas están contenidas en el rectángulo $(\pm\sigma_1, \pm\sigma_2)$, que son tangentes a dicho rectángulo en los puntos $(\sigma_1, \rho\,\sigma_2)$ y $(\rho\,\sigma_1, \sigma_2)$.

(f) Muestre que $f(x_2|x_1)$ es gaussiana, con $N(\mu_2 + \rho(\sigma_2/\sigma_1)(x_1 - \mu_1), \sigma_2\sqrt{1-\rho^2})$. Discuta cómo varía la esperanza de x_2 en función de x_1 según el signo de ρ, y analice cómo varía el ancho de la distribución condicional con el grado de correlación. Interprete estos resultados cortando con líneas x_1=cte las elipses dibujadas a mano alzada en el item anterior. ¿Qué ocurre en el caso límite $\rho = 1$?

8. Aplicando los resultados del ejercicio anterior para el caso de $\underline{x}$ bidimensional,

(a) Dibuje a mano alzada elipses de covarianza con distintos ρ para el caso $\sigma_1 = \sigma_2$. Discuta la diferencia entre tomar como error para X_1 el rango máximo cubierto por la elipse sobre el eje x_1, o el segmento entre los puntos de intersección de la elipse con el eje x_1.

(b) ¿Por qué tiene más sentido considerar la elipse como rango de confianza, que el propio rectángulo $(\pm k\sigma_1, \pm k\sigma_2)$?

(c) Considere las elipses de covarianza encerradas dentro del rectángulo $(\pm k\sigma_1, \pm k\sigma_2)$ alrededor de (μ_1, μ_2). Muestre que la probabilidad conjunta de que (x_1, x_2) se encuentre dentro de una de estas elipses con k=1 es 39.3%, independientemente del valor de la correlación ρ (este resultado es el equivalente al 68.3% obtenido para el caso n=1). Sugerencia: pensar en otro suceso que tenga la *misma* probabilidad que el suceso "(x_1, x_2) se encuentra dentro de una de estas elipses" y que involucre a la variable aleatoria Q.

(d) ¿Cuánto debería ser k para que la elipse corresponda a un nivel de confianza de 95%? Verifique que este resultado puede obtenerse también analíticamente (para el caso bidimensional), además de usando las tablas. [Rta: k=2.448]

9. La función característica $\phi(\underline{t})$ de la distribución multinormal esta dada por

$$\phi(\underline{t}) = \exp\left(i\,\underline{t}^T\underline{\mu} - \frac{1}{2}\,\underline{t}^T\mathbb{V}\,\underline{t}\right)$$

Muestre que $\underline{\mu}$ y $\mathbb{V}$ son en efecto la esperanza y la matriz de covarianza de la variable aleatoria multidimensional $\underline{x}$.

Ayuda: alcanza con mostrar que μ_l es la esperanza de x_l (la componente l-ésima de $\underline{x}$) y que $\mathbb{V}_{kl}$ (el elemento kl de la matriz $\mathbb{V}$) es la covarianza de x_k con x_l.

10. Al realizar mediciones de una variable continua X, se tiene un límite de resolución experimental δ (también conocido como error de cuantización, entre otros nombres). ¿Cómo nos afecta este límite?

Asumamos que la variable continua subyacente X, sin límite de resolución, tiene distribución normal $N(\mu, \sigma^2)$. Por lo tanto, el promedio de n muestras tendrá distribución $N(\mu, \sigma^2/n)$. Llamemos $\widetilde{X}$ a la variable que medimos, limitada por la resolución.

(a) ¿Cómo simularía el efecto del límite de resolución δ? Es decir, ¿cómo obtendría una muestra de $\widetilde{X}$ a partir de una muestra de X?

(b) ¿Cambian la esperanza $E[\widetilde{X}]$, la varianza $\mathrm{Var}(\widetilde{X})$, y la varianza del promedio de n de $\mathrm{Var}(\frac{1}{n}\sum_i^n \widetilde{X}_i)$ respecto de la de X? ¿Dependen de la resolución δ?

Bibliografía:

- Kollar, I. (1994). Bias of mean value and mean square value measurements based on quantized data. IEEE Transactions on Instrumentation and Measurement. https://doi.org/10.1109/19.328894
- Widrow, B., Kollar, I., & Ming-Chang Liu. (1996). Statistical theory of quantization. IEEE Transactions on Instrumentation and Measurement. https://doi.org/10.1109/19.492748