Estadística en Física Experimental (1er cuatrimestre de 2015)

Guía de Problemas N° 4 | Distribuciones multidimensionales – Covarianza y Correlación

1. Demuestre las siguientes propiedades que involucran a la esperanza y/o a la matriz de covarianza de variables aleatorias arbitrarias X, Y.
   (a) $E(aX + bY + c) = a\,E(X) + b\,E(Y) + c$
   (b) $E(E(X)) = E(X)$
   (c) $E(E(X|Y)) = E(X)$, note que una vez evaluada la esperanza condicional $E(X|Y)$ ésta es función de Y
   (d) $E(XY) - E(X)E(Y) = 0$, si X e Y son independientes
   (e) $\mathrm{Cov}(X, Y) = E(XY) - E(X)E(Y)$
   (f) $\mathrm{Var}(X) = E(X^2) - E(X)^2$
   (g) $\mathrm{Var}(aX + bY + c) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y) + 2ab\,\mathrm{Cov}(X, Y)$
   (h) $\mathrm{Var}(a_1X_1 + a_2X_2 + \ldots + a_NX_N) = \sum_{i=1}^{N} a_i^2\mathrm{Var}(X_i) + 2\sum_{i=1}^{N}\sum_{j>i}^{N} a_ia_j\mathrm{Cov}(X_i, X_j)$, que es una generalización del ejercicio anterior (con $c = 0$)

2. Para cada uno de los cuatro pares de datos:

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

   (a) calcular (i) la media muestral de X y de Y, (ii) la varianza muestral de X y de Y, (iii) la correlación entre X e Y: $\rho(X, Y) = \dfrac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}} = \dfrac{\mathrm{Cov}(X, Y)}{\sigma_X\sigma_Y}$;
   (b) grafique cada par de puntos.

   Nota: estos son cuatro conjuntos de datos que F.J. Anscombe generó para mostrar que hacer buenos gráficos de los datos son una parte esencial del análisis de regresión lineal. F. J, Anscombe, (1973). "Graphs in Statistical Analysis". Am Stat, Vol. 27, No. 1, 17-21

3. Sea X una variable aleatoria con densidad de probabilidad simétrica alrededor de cero. Muestre que X e $Y{=}X^2$, pese a no ser independientes, tienen correlación nula.

4. La suma de dos distribuciones uniformes es una distribución triangular:
   (a) Si X e Y son variables independientes con distribución uniforme en [0,1], halle la distribución conjunta $g(U, V)$ de $U \equiv X + Y$ y $V \equiv X - Y$.
   (b) Tomando la correspondiente distribución marginal, muestre que U es una variable aleatoria con distribución triangular:

$$f_U(t) = \begin{cases} t & 0 < t < 1 \\ 2 - t & 1 < t < 2 \\ 0 & \text{en otro caso} \end{cases}$$

   (c) Encuentre la distribución de V y determine si U y V son independientes.
   (d) Calcule la varianza de U via $\displaystyle\int_0^2 (t-1)^2 f_U(t)\mathrm{d}t$. Confirme que se obtiene lo mismo usando la propiedad del ejercicio 1g

5. La suma de gaussianas es gaussiana:
   (a) Probar que si X e Y son variables independientes con distribución normal de parámetros (μ_1, σ_1) y (μ_2, σ_2), entonces $Z = X + Y$ es una gaussiana de parámetros $(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2})$.
   (b) Ahora bien, si la suma de gaussianas es gaussiana, ¿cómo es que en la guía 3 mostró que con la suma de dos gaussianas se consigue aproximar una distribución de Cauchy, que claramente no es gaussiana?

6. Muestre que el cociente $Z \equiv X/Y$ de dos variables independientes con distribución normal canónica tiene distribución de Cauchy, $f_Z(t) = 1/[\pi(1 + t^2)]$.

7. Sea (X, Y) una variable aleatoria bidimensional donde Ω_X=$\{x_1, x_2\}$ e Ω_Y=$\{y_1, y_2\}$, y considere la siguiente asignación de probabilidades: $P(x_1, y_1) = P(x_2, y_2) = p$ y $P(x_1, y_2) = P(x_2, y_1) = q$. Indique qué condiciones deben satisfacer p y q para que:
   (a) P sea una medida de probabilidad.
   (b) X e Y sean independientes.

8. Sean X e Y dos variables independientes con distribución uniforme en [0,1], a partir de las cuales se definen $U = \sqrt{-2a \ln X}\cos(2\pi Y)$ y $V = \sqrt{-2a \ln X}\sin(2\pi Y)$. Encuentre la distribución conjunta $g(U, V)$, identifique qué distribución es, indique el significado del parámetro a y determine si U y V son independientes.

9. *Método de Monte Carlo.* Para generar números pseudoaleatorios con una distribución arbitraria $f(t)$, en un dominio $[a, b]$ en el que f está acotada (siendo f_m su valor máximo) se procede de la siguiente manera:
   I. se generan dos números al azar y y z con distribución uniforme en $[0, 1]$.
   II. A partir de y y z se determinan $u = a + (b - a)y$ y $v = f_m z$.
   III. Si $v \leq f(u)$, se incluye $x = u$ en la muestra de números generados, en otro caso se descarta.
   IV. Se repite el procedimiento hasta obtener la cantidad deseada de números aceptados.

   Para este procedimiento:
   (a) Muestre que u tiene distribución uniforme en $[a, b]$ y v tiene distribución uniforme en $[0, f_m]$.
   (b) Demuestre que la densidad de probabilidad de x es f.
   (c) ¿Es necesario conocer la constante de normalización de f para utilizar este procedimiento?
   (d) Muestre que la fracción de números que se incluye en la muestra, respecto del total de números generados, depende del área bajo la curva de f. ¿Qué concluye acerca de la eficiencia del método?

   Nota: este método propuesto por von Neumann se llama *acceptance-rejection* o también *rejection sampling*. Ver por ejemplo `http://pdg.lbl.gov/2014/reviews/rpp2014-rev-monte-carlo-techniques.pdf`.

10. **Ejercicio para entregar**. *Generación al azar de variables multidimensionales por el método de Monte Carlo.* Se requiere simular numéricamente el comportamiento de una esfera de fluído (inicialmente en reposo) con una densidad
$$\rho(r) = \frac{\rho_0}{1 + \left(\frac{r}{r_e}\right)^2} \quad (0 < r < r_e)$$
   donde ρ_0 es la densidad central y r_e el radio de la esfera. Para ello se generan como condiciones iniciales las posiciones $\overrightarrow{r}$ de 10000 elementos de fluído con una distribución proporcional a la masa (i.e., la probabilidad de encontrar un elemento de fluído en un determinado dV alrededor de $\overrightarrow{r}$ es proporcional a ρdV), y luego se calcula la evolución temporal de las mismas integrando las ecuaciones de la hidrodinámica.Utilice el método descripto en el ejercicio 9 para generar las condiciones iniciales del experimento en coordenadas esféricas (considere $r_e = 1$). Para ello:
   (a) Escriba la densidad de probabilidad conjunta de r, θ y φ (las coordenadas esféricas usuales).
   (b) Calcule las distribuciones de probabilidad marginales de cada una de las coordenadas por separado. ¿Son estas independientes?
   (c) Genere 10000 valores al azar de cada variable, usando el método de Monte Carlo. Grafique los histogramas de las distintas variables, y superponga las distribuciones correspondientes.

11. 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 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) Considere m variables aleatorias $\underline{y}$, obtenidas como función lineal $\underline{y} = \mathbb{C}\underline{x}$ a partir de n variables $\underline{x}$ con distribución conjunta multinormal, siendo $\mathbb{C}$ de $m \times n$, con $m < n$. Muestre que $\underline{y}$ tiene también distribución multinormal con covarianza $\mathbb{V}' = \mathbb{C}\mathbb{V}\mathbb{C}^T$.

(c) Sea $\mathbb{S}$ la matriz ortogonal que diagonaliza $\mathbb{V}$ para n arbitrario. Muestre que la distribución de $\underline{z} = \mathbb{S}\underline{x}$ corresponde a n variables gaussianas independientes. A partir de ésto encuentre la distribución de la variable aleatoria $Q = (\underline{x}-\underline{\mu})^T \mathbb{V}^{-1}(\underline{x}-\underline{\mu})$ que aparece en el exponente de la multinormal.

12. *Multinormal bivariada, elipses de covarianza.*

(a) En el caso $n = 2$ la matriz de covarianza de una multinormal depende de 3 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

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

(b) 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.

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

(d) 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 $\rho = 1$?

(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 $(\rho\,\sigma_1, \sigma_2)$ y que su eje principal forma un ángulo ϕ con el eje x_1 dado por $\tan 2\phi = 2\rho\,\sigma_1\sigma_2\,/(\sigma_1^2 - \sigma_2^2)$.

(f) 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.

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

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

(a) 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 equivalentente 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)

(b) ¿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]

Histogramas

14. Un histograma es una representación gráfica de un conjunto de datos (simulados o reales) que permite estimar la densidad de probabilidad de una variable contínua. Si estamos interesados en lo que ocurre con un solo

'bin', indique qué distribución de probabilidad se puede usar para describir la variable aleatoria "número de datos que caen en el i-ésimo bin". ¿Cuál es la varianza estimada de la altura del bin? ¿Que error incluiría en el gráfico?

15. Consideremos n datos y un histograma con un número k de bines. Podemos pensar a este histograma como una manera de distribuir los n datos en k cajitas, siguiendo una probabilidad conjunta. La distribución multinomial, que es una generalización de la binomial, en la que no existen solo dos clases de resultados ("éxito" y "fracaso") sino k posibles resultados está dada por:

$$M(r_1, \ldots, r_k; n, p_1, \ldots, p_k) = \frac{n!}{r_1! \ldots r_k!} {p_1}^{r_1} \ldots {p_k}^{r_k}$$

con estos dos vínculos: $\sum_{i=0}^{k} p_i = 1$ y $\sum_{i=0}^{k} r_i = n$. Suponiendo conocidas las varianzas y las covarianzas del número de entradas en cada bin (ver Frodesen pg 72): muestre que las entradas de distintos bines tienden a estar descorrelacionadas si las probabilidades de cada uno son muy chicas. En este límite ¿son independientes el número de entradas en distintos bines?

16. Consideremos ahora que el número de datos medidos es aleatorio con distribución de Poisson: $P(n; \lambda)$. Muestre que la variable aleatoria el "número de datos que caen en el i-ésimo bin" tiene distribución de Poisson $P(r_i; \lambda p_i)$, con p_i la probabilidad de que un dato caiga en el i-ésimo bin. ¿Son independientes las entradas de distintos bines? ¿Qué error reportaría para cada bin en este caso?

17. ¿Como ajustaría la altura del 'bin', si el ancho de los bines no es uniforme a lo largo del eje horizontal?