Máster Universitario en Ciencia de Datos
Fundamentos para el Análisis de Datos y la Investigación.
Prácticas Temas 5, 6 y 7. Curso 2023-24. Prof. JM Sarabia

1. Tema 5: Modelos de distribuciones de datos y simulación

1.1. Modelos de conteo y continuos

1. La probabilidad de que un director de una sucursal venda una nueva modalidad de fondo durante una entrevista a un cliente es del 15 por ciento.

   a) Si realiza un total de 25 entrevistas, obtener la distribución de probabilidad del número fondos vendidos. Representar gráficamente la función de densidad y la función de distribución

   b) Obtener la media y la desviación típica del número de fondos vendidos.

   c) Probabilidad de que en cinco visitas realice al menos una venta.

   d) Probabilidad de que en diez visitas realice al menos cuatro ventas.

2. Un examen de estadística consiste en un test de 10 preguntas de respuesta múltiple, con cuatro posibles respuestas por pregunta. Un estudiante completa el examen eligiendo al azar la respuesta de cada pregunta.

   a) Identificar la distribución del número de aciertos.

   b) Tabular las probabilidades de acierto y representarlas gráficamente. Obtener la tabla de las probabilidades acumuladas y representarla gráficamente.

   c) Probabilidad que acierte 5 ó más preguntas. Probabilidad que acierte 5 preguntas ó menos.

3. El número de errores cometidos por un programador en una jornada diaria sigue una distribución de Poisson de media 4.

   a) Probabilidad de que en una jornada cometa al menos un fallo.

   b) Obtener las probabilidades de que en una jornada cometa $x = 0, 1, 2, 3, 4$ y fallos.

   c) Probabilidad de que en una jornada cometa x fallos o menos, para $x = 0, 1, \ldots, 5$.

   d) Representar las probabilidades brutas y acumuladas de x fallos, para valores de $x = 0, 1, \ldots, 12$.

4. El número de contactos de los usuarios de una red social sigue una distribución de Poisson de media 75 contactos. Simular una muestra del número de contactos de $n = 500$ usuarios de la red.

   a) Obtener las características de la muestra simulada: mínimo, máximo, media, etc.

   b) Número esperado de usuarios con más de 90 contactos.

   c) Representar gráficamente la muestra y la frecuencia esperada teórica según el modelo de Poisson. ¿Qué se observa?

5. El tiempo de retraso en minutos de un vuelo sigue una distribución uniforme continua entre 1 y 7 minutos.

a) Representar gráficamente la función de densidad y de distribución del tiempo de retraso.

b) Simular una muestra de n=10.000 tiempos de retraso. Representar el histograma de la muestra y una estimación de la densidad. Obtener las medias y desviaciones típicas teóricas y muestrales.

6. El tiempo de servicio en una sucursal bancaria sigue una distribucuón exponencial de media 3 minutos por cliente.

   a) Probabilidad que el tiempo de servicio a un cliente sea inferior a 2 minutos.

   b) Representar las funciones de densidad y de distribución

   c) Simular una muestra del tiempo de espera de 100 clientes. Representar el histograma y la función de densidad.

7. La ley de Benford o ley del primer dígito x establece que la distribución de probabilidad del primer dígito de ciertos conjuntos de datos es:

$$P(X = x) = \log_{10}(1 + 1/x), \;\; x = 1, 2, \ldots, 9$$

   donde $\log_{10}(x)$ es el logaritmo decimal. Esta ley es utilizada para la detección del fraude en diversos contextos.

   a) Tabular las probabilidades de la ley de Benford, directamente y a través de su función en `R`.

   b) Como aplicación consideramos la tabla de la distribución del primer dígito de los municipios de la Comunidad de Navarra

Dígito	1	2	3	4	5	6	7	8	9
Frecuencia	91	54	37	21	18	16	13	15	8

   Ajustar una ley de Benford a los datos y validarla mediante el test de la chi-cuadrado de Pearson. Repetir el ejercicio con los municipios de España (8.109)

Dígito	1	2	3	4	5	6	7	8	9
Frecuencia	2559	1503	992	731	635	540	456	358	335

8. Usando el método de la transformación inversa, generar una muestra aleatoria de tamaño n=1000 de la distribución exponencial con media 5, y función de distribución $F(x) = 1 - \exp(-x/5)$, si $x > 0$.

9. Por medio de la transformación de Box-Muller, generar una muestra de 100 datos de una distribución normal de media 1 y desviación típica 0.5. Validar los resultados gráfica y analíticamenre mediante con el contraste de Shapiro-Wilk.

1.2. Modelos de dos o más dimensiones

1. Haciendo uso del dataset `Swiss`, se trata de obtener una muestra simulada de tres variables correlacionadas del dataset. Para ello, elegir tres variables cuya distribución sea normal, obtener la matriz de covarianzas, y simular 100×3 observaciones. Hacer uso de la descomposición de Cholesky.

2. Simular una distribución normal bidimensional (X, Y) con parámetros $\mu_1 = \mu_2 = 0$, $\sigma_1 = \sigma_2 = 1$ y $\rho = 0,8$, mediante la simulación de la distribución marginal X y de la distribución condicionada $Y|X$.

3. Los precios de dos acciones de bolsa se pueden modelizar según una distribución normal bidimensional, con vector de medias $(2; 2{,}5)$, $\sigma_X = 0{,}55$, $\sigma_Y = 0{,}45$ y coeficiente de correlación lineal $\rho_{XY} = 0{,}71$

   (i) Representar la función de densidad mediante el comando `persp`, definiendo previamente la función de densidad. Representar los contornos.

   (ii) Haciendo uso del comando `scatterplot3d` representar la función de densidad.

4. Los rendimientos de dos activos de bolsa se pueden modelizar mediante una variable normal bidimensional $(X_1, X_2)^\top$, donde $\mu_1 = 1$, $\mu_2 = 0,5$, $\sigma_1 = 0,55$, $\sigma_2 = 0,7$ y $\sigma_{12} = 0,3$.

   (i) Media y varianza de $0,7X_1 + 0,3X_2$

   (ii) Hallar $P(0,7 \cdot X_1 + 0,3 \cdot X_2 \leq 1,8)$ de forma exacta y mediante simulación.

5. La variable normal trivariada $(X_1, X_2, X_3)^\top$ representa el los rendimientos de tres activos en un mercado, de modo que el vector de medias es $(0, 0, 0)$ y matriz de covarianzas $\sigma_{ii} = 1$, $\sigma_{12} = 3/5$, $\sigma_{13} = 1/3$ y $\sigma_{23} = 11/15$. Se trata de calcular la probabilidad:

$$P(X_1 < 1,\ X_2 < 4,\ X_3 < 2)$$

   a) De forma exacta, mediante el paquete `mvtnorm`.

   b) Mediante simulación.

6. En una junta de accionistas, el 55 por ciento de los delegados llega en avión, el 25 por ciento en tren y el resto en coche. Se selecciona una muestra de 8 delegados.

   a) Probabilidad que 5 hayan llegado en avión, 2 en tren y 1 en coche

   b) Probabilidad de que todos hayan llegado en avión

   c) Probablidad que 4 lleguen en avión y 4 en tren

   d) Simular lo que ocurre en 6 juntas de accionistas suponiendo que a cada junta acuden 25 delegados.

7. Se lanza un dado n veces, y se calcula la media de los resultados obtenidos.

   a) ¿A qué valor converge (en probabilidad) la media muestral cuando $n \to \infty$?

   b) Haciendo uso de `R`, probar el resultado anterior.

8. Un científico de datos quiere obtener la distribución de las ventas medias mensuales de una app de ventas online mediante simulación. Conoce que las ventas en un día siguen una distribución con media 150 euros y desviación típica 12

   a) Obtener mediante simulación la distribución de la media muestral en 30 días

   b) Hallar mediante `replicate` la distribución en el muestreo de la media muestral con $m = 10^4$

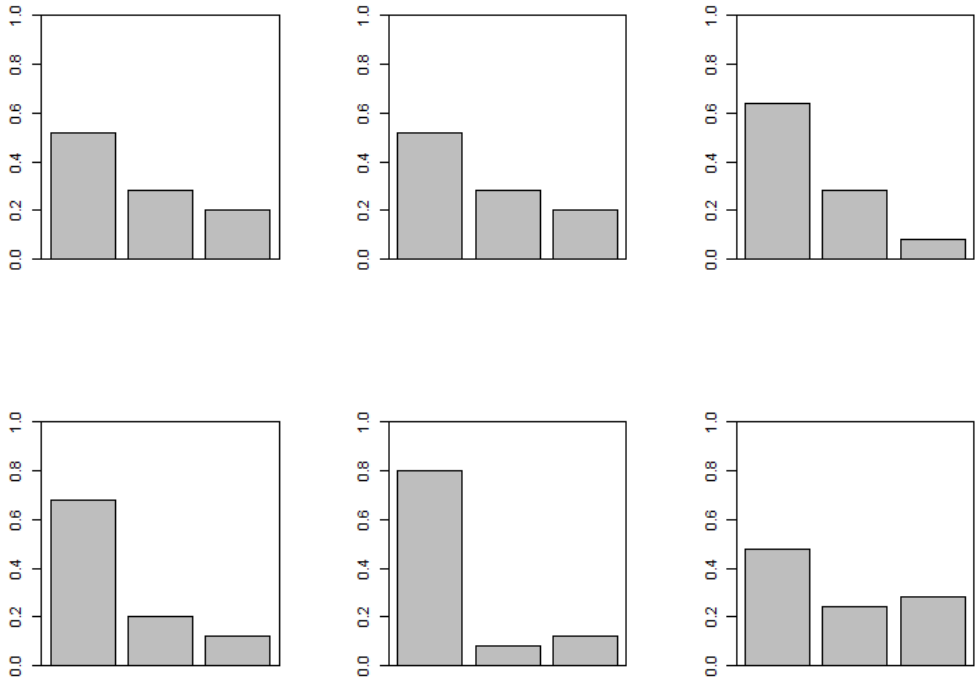

Figura 1: Distribución de llegadas en 6 juntas

2. Tema 6: Experimentos con datos y pruebas estadísticas significativas

2.1. Pruebas A/B y de diferencias

1. Una consultora quiere estudiar la diferencia entre dos páginas web la A y la B, de complejidad similar, de modo que las diferencias radican en la presentación y diseño de cada página. Para ello elige 12 clientes y asigna al azar 6 a cada una de las presentaciones. A continuación mide el tiempo en segundos en que cada cliente consulta la página.

Cliente N	Tiempo Web	Grupo
1	58	A
2	89	A
3	57	B
4	81	B
5	50	A
6	62	A
7	22	B
8	49	A
9	56	A
10	43	B
11	67	B
12	60	B

   Estudiar gráfica y analíticamente si existen diferencias entre las dos presentaciones web mediante pruebas significativas.

2. Una empresa trata de vender un servicio a través de una página web. Para ello realiza un experimento y considera dos páginas web con dos precios A y B, registrando los datos sin y con coversión según el tipo de web:

Resultado	Precio A	Precio B
Conversión	200	182
Sin conversión	23539	22406

   Probar que el precio A convierte cerca de un 5 % mejor que el precio B. Se trata de estudiar si existen diferencias significativas en los dos precios. Hacer uso de la función `prop.test`

3. Utilizando el fichero de datos `mtcars`, comparar la variable `mpg` en los coches automáticos y de marchas.

   a) Contrastar la normalidad de las dos muestras mediante el contraste de Shapiro-Wilk.

   b) Obtener un intervalo de confianza para la diferencia de `mpg`, con un nivel de confianza del 95 por ciento. ¿Existen diferencias significativas?

   c) Obtener un intervalo de confianza para el cociente de varianzas. ¿Se puede concluir que las varianzas de las dos variables son iguales? Obtener un nuevo intervalo de confianza para la diferencia de medias, teniendo en cuenta la hipótesis de igualdad de varianzas

4. El fichero de datos `immer` incluye información sobre rendimientos de variedades de cebada en dos años. Se trata de comparar los rendimientos en las dos fechas.

   a) Contrastar la hipótesis de normalidad de la diferencia de rendimientos.

   b) Hallar un intervalo de confianza para la media y la desviación típica de los rendimientos.

   c) Contrastar la hipótesis de que la diferencia de los rendimientos medios en las dos fechas.

5. En las elecciones a representante sindical de una empresa con 600 trabajadores, 270 empleados se muestran a favor de un cambio en la dirección del sindicato.

   a) Hallar intervalos de confianza para la verdadera proporción de trabajadores a favor del cambio en la dirección con niveles de confianza del 95 y 99 por ciento.

   b) Contrastar la hipótesis $H_0 : p = 0,5$ frente a $H_1 : p \neq 0,5$.

   c) En una segunda encuesta a 220 trabajadores, 94 se muestran a favor. Obtener el intervalo de confianza para la diferencia de proporciones, y contrastar la hipótesis de igualdad de proporciones.

   d) Número de encuestas que hay que realizar para obtener un error máximo de estimación del 6 por ciento con un nivel de confianza del 95 %. Construir una función para obtener el número de encuestas. Realizar un grafico que represente el tamaño muestral según diferentes valores del error.

2.2. Pruebas chi-cuadrado y de bondad de ajuste

1. Estamos probando tres titulares diferentes A, B y C y los ejecutamos cada uno en 1000 viistantes, obteniéndose la siguiente tabla:

	Titular A	Titular B	Titular C
Clic	14	8	12
Sin clic	986	992	988

   Queremos estudiar si exiten diferencias en los tres titulares

   a) Obtener la tabla, suponiendo que los tres titulares tienen la misma tasa de clics

   b) Contrastes la hipótesis de igualdad de clics, mediante el test de la chi-cuadrado exacto y mediante simulación

   c) Hacer uso de la prueba exacta de Fisher

2. Un analista clasifica los fondos de inversión en cuatro categorías según el riesgo: bajo, medio, alto y muy alto en proporciones 0.45; 0.35; 0.15 y 0.05. Se elige una muestra de 100 fondos y un grupo de expertos los clasifica en la escala anterior con frecuencias 49, 27, 14 y 10. ¿Se ajustan las proporciones propuestas por el analista a la clasificación de los expertos?

3. Una entidad financiera está reestructurando el número de oficinas en una zona. En una sucursal de la zona, durante un total de 100 días elegidos al azar se ha registrado la variable número diario de cancelaciones de cuentas, obteniéndose la siguiente tabla,

número de cancelaciones	0	1	2	3	4+
frecuencia	32	37	20	7	4

Contrastar la hipótesis que el número de cancelaciones sigue una distribución de Poisson de parámetro 1.1

4. Contrastar la hipótesis de que un dado está equilibrado, si en 100 lanzamientos se han obtenidos las frecuencias de: 15, 18, 14, 20, 12 y 21.

5. Los siguientes datos representan el tiempo de servicio en una entidad bancaria

   0,4 3,8 3,9 2,2 4,9 12,6 7,9 15,3 11,5 6,3

   Contratar la hipótesis que la distribución de los tiempos sigue una distribución exponencial de media 5 minutos, mediante el contraste de Kolmogorov-Smirnov.

6. Para la base de datos Wage (paquete ISLR), se trata de contrastar la hipótesis de homogeneidad/independencia entre algunas variables categóricas.

   a) Contruir la tabla de contingencia entre las variables race y education. Obtener las distribuciones marginales a partir de la tabla. Contrastar la hipótesis de independencia. Análizar en qué casillas se producen las mayores discrepancias en relación con la hipótesis de independencia.

   b) Realizar el mismo estudio con las variables health y education y race y health.

7. La siguiente tabla relaciona número de fumadores en cuatro filiales de una empresa:

Filial \fumadores	fumadores	no fumadores
F_1	14	88
F_2	17	83
F_3	18	82
F_3	16	80

   ¿Se pueden considerar las filiales homogéneas en lo relativo al número de fumadores?

8. Contrastar la hipótesis de aleatoriedad mediante el test de Wald-Wolfowitz de las variables de las bases de datos:

   a) `swiss`

   b) `Cotizaciones2020.txt`, sobre precios y rendimientos

2.3. Pruebas múltiples A/B/...

1. Con los datos de adherencia de las cuatro páginas web:

   a) Estudiar si existen diferencias signiifcativas entre los tiempos medios de las cuatro páginas

   b) Contrastar las hipótesis del modelo ANOVA

   c) Estudiar si existen diferencias signifcativas entre las páginas dos a dos

2. Cuarenta vendedores de una empresa reciben cursos de formación sobre técnicas de ventas. Existen 4 programas de formación, a los que se asignan 10 empleados por programa. Se desea estudiar las posibles diferencias entre los cuatro programas. Una vez finalizados los programas, y al cabo de un cierto tiempo, se registra el número de unidades vendidas de un determinado producto.

Programa A	Programa B	Programa C	Programa D
95	82	85	78
88	88	81	65
90	80	86	74
99	75	91	82
89	67	78	75
93	78	81	62
95	81	86	75
97	80	90	79
85	77	75	70
90	69	83	82

   a) Contrastar la hipótesis de que el número de unidades vendidas sigue una distribución normal. Hallar el número medio de unidades vendidas por programa, junto con los correspondientes intervalos de confianza.

   b) ¿Existen diferencias significativas entre las ventas medias de los cuatro programas? Obtener la tabla ANOVA e interpretar sus elementos.

   c) Contrastar la hipótesis de que el número de unidades vendidas no depende del programa, mediante el contraste de Kruskal Wallis.

3. Tema 7: Cadenas de Markov

1. El mercado de telefonía móvil presenta cambios importantes en periodos relativamente cortos. Supongamos las siguientes probabilidades de transición entre las tres compañías líderes de este año para el siguiente:

$i \to j$	Movistar	Orange	Vodafone
Movistar	0.37	0.18	0.45
Orange	0.23	0.45	0.32
Vodafone	0.22	0.30	0.48

   a) Obtener el diagrama de la matriz de transición y la clasificación de estados

   b) Partiendo de la cuota de mercado actual $p^{(0)} = (\frac{23}{78}, \frac{30}{78}, \frac{25}{78})$, obtener las cuotas de mercado despúes de uno, dos y tres años

   c) Hallar la Distribución estacionaria