

Diagrama Chi cuadrado

- Existe un método más formal para evaluar la normalidad conjunta para un conjunto de datos, basado en la distancia cuadrada generalizada

$$d_j^2 = (\mathbf{x}_j - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_j - \bar{\mathbf{x}}), j = 1, 2, \dots, n$$

donde $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ son muestras de observaciones.

- Cuando la población raíz es normal multivariada y n y $n-p$ son más grandes que 25 o 30, cada una de las distancias cuadradas $d_1^2, d_2^2, \dots, d_n^2$ deben comportarse como una chi-cuadrada.
- La idea es graficar los cuantiles de las distancias de la muestra vs los cuantiles esperados de una distribución chi cuadrada.
- Si el diagrama se parece a una línea recta de 45 grados que pasa por el origen, esto sugiere que los datos son normales.

Pasos para construir un diagrama Chi-cuadrado:

- 1 Se ordenan las distancias cuadradas de la más pequeña a la más grande como $d_{(1)}^2 \leq d_{(2)}^2 \leq \dots \leq d_{(n)}^2$
- 2 Se encuentran los cuantiles correspondientes, $q_{c,p}(\frac{j-0.5}{n})$, de la distribución chi cuadrada con p grados de libertad.
- 3 Se grafican los pares de cuantiles $(q_{c,p}(\frac{j-0.5}{n}), d_j^2)$. Si los puntos están sobre o cerca de una línea recta de 45 grados, esto apoya la suposición de que las observaciones son elegidas de una población normal.

Ejemplo: Construcción del diagrama chi-cuadrado para la muestra de datos bivariados

- Para nuestros datos bivariados tenemos

$$\bar{\mathbf{x}} = \begin{bmatrix} 5.26 \\ -3.09 \end{bmatrix}$$
$$\mathbf{S} = \begin{bmatrix} 7.12 & -0.67 \\ -0.67 & 12.43 \end{bmatrix} \rightarrow \mathbf{S}^{-1} = \begin{bmatrix} 0.1411 & 0.0076 \\ 0.0076 & 0.0808 \end{bmatrix}$$

Ejemplo: Construcción del diagrama chi-cuadrado para la muestra de datos bivariados

- Las distancias cuadradas generalizadas de cada punto al centroide \bar{x} están dadas en la tabla de la izquierda

x_{j1}	x_{j2}	d_j^2
1.43	-0.69	2.400
1.62	-5.00	2.279
2.46	-1.13	1.336
2.48	-5.20	1.548
2.97	-6.39	1.739
4.03	2.87	2.976
4.47	-7.88	2.005
5.76	-3.97	0.090
6.61	2.32	2.737
6.68	-3.24	0.281
6.79	-3.56	0.333
7.46	1.61	2.622
7.88	-1.87	1.138
8.92	-6.60	2.686
9.42	-7.64	3.819

x_{j1}	x_{j2}	d_j^2
5.76	-3.97	0.090
6.68	-3.24	0.281
6.79	-3.56	0.333
7.88	-1.87	1.138
2.46	-1.13	1.336
2.48	-5.20	1.548
2.97	-6.39	1.739
4.47	-7.88	2.005
1.62	-5.00	2.279
1.43	-0.69	2.400
7.46	1.61	2.622
8.92	-6.60	2.686
6.61	2.32	2.737
4.03	2.87	2.976
9.42	-7.64	3.819

Ordenamos
las distancias
cuadradas
generalizadas
(de menor a
mayor)

Ejemplo: Construcción del diagrama chi-cuadrado para la muestra de datos bivariados

Entonces se encuentran los correspondientes percentiles

$100 \left(\frac{j-0.5}{n} \right)^{th}$ de la distribución Chi-cuadrada con 2 grados de libertad

x_{j1}	x_{j2}	d_j^2	$(j-0.5)/n$	$q_{c,2}[(j-0.5)/n]$
5.76	-3.97	0.090	0.033	0.068
6.68	-3.24	0.281	0.100	0.211
6.79	-3.56	0.333	0.167	0.365
7.88	-1.87	1.138	0.233	0.531
2.46	-1.13	1.336	0.300	0.713
2.48	-5.20	1.548	0.367	0.914
2.97	-6.39	1.739	0.433	1.136
4.47	-7.88	2.005	0.500	1.386
1.62	-5.00	2.279	0.567	1.672
1.43	-0.69	2.400	0.633	2.007
7.46	1.61	2.622	0.700	2.408
8.92	-6.60	2.686	0.767	2.911
6.61	2.32	2.737	0.833	3.584
4.03	2.87	2.976	0.900	4.605
9.42	-7.64	3.819	0.967	6.802

Construimos ahora el scatter plot de los pares

$$\left(q_{c,p}\left(\frac{j-0.5}{n}\right), d_j^2 \right)$$

Si estos puntos están sobre una línea

recta, los datos apoyan la suposición

Ejemplo: Construcción del diagrama chi-cuadrado para la muestra de datos bivariado

La gráfica no parece apoyar la suposición de que las observaciones fueron extraídos de una población normal bivariada

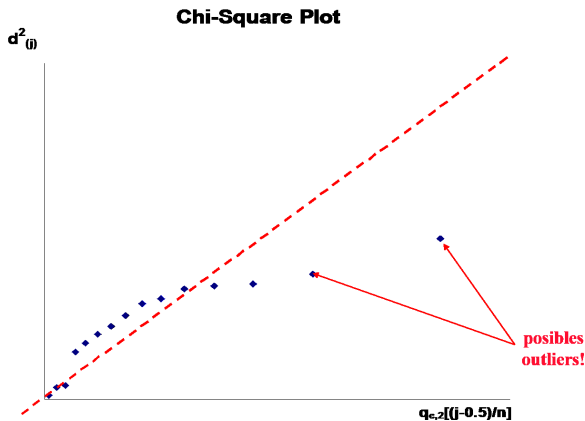




Diagrama chi-cuadrado

- Este procedimiento no está limitado al caso bivariado; puede usarse para evaluar normalidad en dimensiones mayores a 2 ($p > 2$). 
- Algunos investigadores también sugieren calcular la correlación entre $d_{(j)}$ y $q_{c,p}[(j - .5)/n]$ y usar la prueba de Looney & Gullledge  para evaluar la rectitud del diagrama chi-cuadrado.
- Para el ejemplo la correlación es $r_Q = 0.8952$.
- De la tabla para la prueba de normalidad basada en el coeficiente de correlación, y considerando un tamaño de muestra $n = 15$, los puntos críticos son 0.9503 en $\alpha = 0.10$, 0.9389 en $\alpha = 0.05$, y 0.9126 en $\alpha = 0.01$.
- Por tanto rechazamos la hipótesis de normalidad en cualquier α mas grande que 0.01

Resumen de las técnicas para evaluar la suposición de normalidad multivariada.

Se basan principalmente en el cálculo de

$d_j^2 = (\mathbf{x}_j - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_j - \bar{\mathbf{x}})$, $j = 1, 2, \dots, n$, y éstas se comparan con los cuantiles de la χ^2 . La normalidad p -variada es adecuada si

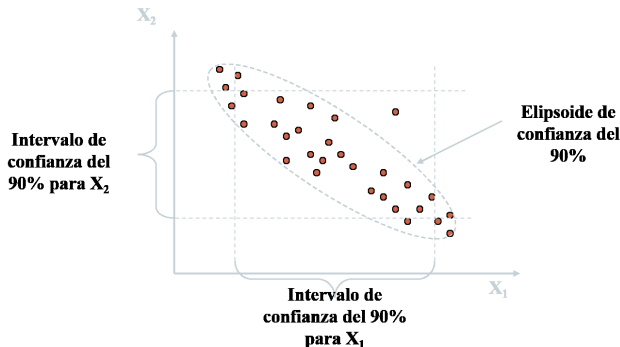
- 1 Aproximadamente la mitad de las d_j^2 son menores o iguales al cuantil del 50%, $q_{c,p}(0.5)$, de la distribución chi-cuadrada con p grados de libertad
- 2 O bien, si un diagrama de las distancias ordenadas

$$d_{(1)}^2 \leq d_{(2)}^2 \leq \dots \leq d_{(n)}^2 \quad \text{vs} \quad q_{c,p}\left(\frac{1-0.5}{n}\right), q_{c,p}\left(\frac{2-0.5}{n}\right), \dots, q_{c,p}\left(\frac{n-0.5}{n}\right)$$


respectivos, se aproxima a una línea recta de 45 grados que pasa por el origen.

Detección de Outliers

- Por tanto, este outlier bivariado no puede ser detectado al inspeccionar gráficamente las distribuciones marginales de X_1, X_2 .
- En general, en altas dimensiones pueden existir outliers que no pueden detectarse mediante diagramas univariados e incluso de los scatter plots.



Una estrategia para la detección de outlier multivariados:

- Buscar los outliers univariados
 - Se calculan los valores estandarizados $z_{jk} = (x_{jk} - \bar{x}_k) / \sqrt{s_{kk}}$ para $j = 1, \dots, n$, y cada columna $k = 1, \dots, p$. Se examinan estos valores estandarizados para detectar valores grandes o pequeños.
 - Dot plots, histogramas y Q-Q Plots para cada variable
- Buscar los outliers bivariados
 - Distancias cuadradas generalizadas de cada par de variables a su centroide, analizando estas distancias se podría identificar valores outliers 
 - Scatter plots para cada par de variables.
 - Diagramas Chi-Cuadrado para cada par de variables
- Buscar los outliers p-dimensionales
 - Distancias cuadradas generalizadas
 - Diagramas Chi- cuadrados

Sin embargo NINGUNA ESTRATEGIA garantiza la detección de outliers multivariados!

Ejemplo de detección de outliers

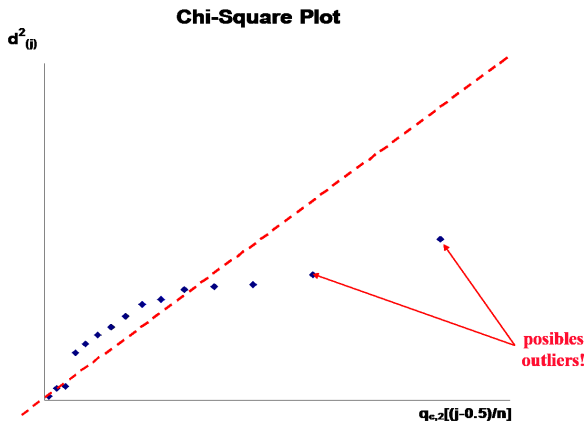
Aquí están calculados los valores estandarizados z_{jk} y las distancias cuadradas generalizadas d_j^2 para los datos anteriores:

x_{j1}	z_{j1}	x_{j2}	z_{j2}	d_j^2
5.76	0.185	-3.97	-0.250	0.090
6.68	0.530	-3.24	-0.043	0.281
6.79	0.570	-3.56	-0.131	0.333
7.88	0.981	-1.87	0.347	1.138
2.46	-1.050	-1.13	0.556	1.336
2.48	-1.045	-5.20	-0.599	1.548
2.97	-0.860	-6.39	-0.936	1.739
4.47	-0.299	-7.88	-1.359	2.005
1.62	-1.367	-5.00	-0.541	2.279
1.43	-1.436	-0.69	0.681	2.400
7.46	0.822	1.61	1.333	2.622
8.92	1.371	-6.60	-0.994	2.686
6.61	0.504	2.32	1.536	2.737
4.03	-0.461	2.87	1.691	2.976
9.42	1.556	-7.64	-1.291	3.819

Esto se ve un poco inusual en
el espacio de $p = 2$

Ejemplo de detección de outliers

Del diagrama Chi-cuadrada se comprueba que este dato está muy alejado de los restantes y del origen, por tanto podría ser un dato outlier



Comentarios finales sobre los outliers

- Si se identifican outliers en un conjunto de datos multivariados, se debe verificar primero su contenido o veracidad.
- Dependiendo de la naturaleza de los outliers y los objetivos de la investigación, éstos pueden eliminarse o ponderarse de una forma apropiada en análisis posteriores.
- Aún cuando muchas técnicas estadísticas asumen poblaciones normales, aquellas que se basan en los vectores de medias muestrales usualmente no serán afectados por un número pequeño de outliers.

Transformaciones para aproximar a normalidad

Que pasa cuando la suposición de normalidad no se cumple en un conjunto de datos?

En este caso existen dos alternativas:

- 1 Ignorar la comprobación de normalidad y proceder como si los datos fueran normalmente distribuidos. Esta práctica no es recomendable, ya que en muchas situaciones esto podría conducir a conclusiones erróneas.
- 2 Convertir los datos no normales a datos que sean aproximadamente normales, mediante transformaciones. Entonces el análisis de la teoría normal se puede realizar sobre los datos transformados.

Transformaciones para aproximar a normalidad

- Una transformación es una reexpresión de los datos en diferentes unidades
- Por ejemplo, cuando un histograma de observaciones presenta una cola larga a la derecha, al transformar las observaciones aplicando logaritmos o raíces cuadradas, frecuentemente se mejora la simetría alrededor de la media y por tanto la aproximación a una distribución normal
- En muchas situaciones, las transformaciones de los datos, proveen expresiones más naturales de las características que se están estudiando

Transformaciones para aproximar a normalidad

- Transformaciones para convertir datos no normales a datos aproximadamente normales se sugieren usualmente por:
 - 1 Consideraciones teóricas
 - 2 Los datos brutos así lo requieren
- Transformaciones útiles para aproximar a la normalidad

Escala Original	Escala transformada
Conteos, y	\sqrt{y}
Proporciones, \hat{p}	$\text{logit}(\hat{p}) = \frac{1}{2} \log \left(\frac{\hat{p}}{1-\hat{p}} \right)$
Correlaciones, r	Transformación z de Fisher, $z(r) = \frac{1}{2} \log \left(\frac{1+r}{1-r} \right)$

Transformaciones de potencia

- En muchas situaciones, la elección de la transformación no es obvia. En este caso, es mejor dejar que los datos sugieran una transformación adecuada.
- Una familia de transformaciones útiles para este propósito son las transformaciones de potencia.
- Para seleccionar una transformación de potencia, el investigador debe analizar en la marginal, los diagramas de puntos o los histogramas y decidir si valores grandes deben ser considerados o descartados para mejorar la simetría alrededor de la media.
- Es común realizar cálculos de prueba y error con algunos valores de las potencias en las transformaciones, hasta obtener una mejor aproximación a la normal.
- La transformación final elegida para los datos debe ser examinada mediante Q-Q plots o cualquier otro criterio, para ver si la suposición de normalidad se cumple.

Transformaciones de potencia para variables continuas

- **Box y Cox [1964]** sugieren un método para encontrar una transformación apropiada a partir de la familia de transformaciones de potencia dada por

$$x^{(\lambda)} = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \ln(x) & \lambda = 0 \end{cases}$$

La cual es continua en λ para $x > 0$.

- Las transformaciones de potencias están definidas únicamente para variables positivas, aunque esto no es muy restrictivo como parece.
- Se le puede sumar una constante m a cada observación en el conjunto de datos, de tal forma que si algunos de los valores son negativos, entonces $x + m > 0$
- Cuando $\lambda > 1$, la transformación produce una mayor separación o dispersión de los valores grandes de x .
- Cuando $\lambda < 1$, el efecto es contrario, valores grandes de x tienden a concentrarse y los valores pequeños a dispersarse

Transformaciones de potencia para variables continuas

Dada las observaciones x_1, \dots, x_n , la solución de Box-Cox para elegir una potencia apropiada λ , es la solución que maximiza la expresión

$$t(\lambda) = -\frac{n}{2} \ln \left[\frac{1}{n} \sum_{j=1}^n (x_j^{(\lambda)} - \bar{x}^{(\lambda)})^2 \right] + (\lambda - 1) \sum_{j=1}^n \ln(x_j)$$

donde

$$x_j^{(\lambda)} = \begin{cases} \frac{x_j^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \ln(x) & \lambda = 0 \end{cases}$$

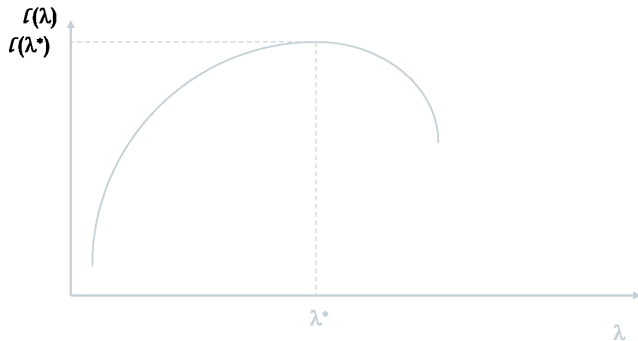
$$\bar{x}^{(\lambda)} = \frac{1}{n} \sum_{j=1}^n x_j^{(\lambda)} = \frac{1}{n} \sum_{j=1}^n \left(\frac{x_j^\lambda - 1}{\lambda} \right)$$

Transformación de Box-Cox

- λ es el parámetro de la transformación que se estima a partir de los datos
- El primer término de $\iota(\lambda)$, aparte de una constante, es el logaritmo de una función de verosimilitud normal, después de maximizarla con respecto a los parámetros poblacionales μ y σ .
- Al encontrar un valor de λ que maximice la función de verosimilitud normal, estamos encontrando una transformación de los datos, que será más aproximada a la distribución normal.

Transformación de Box-Cox

Entonces evaluamos $\iota(\lambda)$ en muchos puntos en un intervalo corto (por decir $[-1,1]$ o $[-2,2]$), graficamos los pares $(\lambda, \iota(\lambda))$ y buscamos el punto máximo



- Frecuentemente en la práctica se elige un valor sencillo de λ cercano a λ^* , como el valor adecuado de la transformación

Transformaciones de Box-Cox de observaciones multivariadas

- Cuando tenemos un conjunto de observaciones multivariadas $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$, se selecciona una transformación de potencia para cada variable. Sean $\lambda_1, \dots, \lambda_p$ las transformaciones de potencias para las p variables. Cada λ_k seleccionada, maximiza

$$t_k(\lambda) = -\frac{n}{2} \ln \left[\frac{1}{n} \sum_{j=1}^n (x_{jk}^{(\lambda_k)} - \bar{x}_k^{(\lambda_k)})^2 \right] + (\lambda_k - 1) \sum_{j=1}^n \ln(x_{jk})$$

- Este procedimiento es equivalente transformar cada distribución marginal a una distribución aproximadamente normal.
- Aunque el hecho de que las marginales sean normales no es suficiente para asegurar que la distribución conjunta sea normal.
- Sin embargo en aplicaciones prácticas esto podría ser suficientemente para asegurar normalidad conjunta.

Transformaciones Box-Cox de observaciones multivariadas

- Para obtener normalidad conjunta, podemos empezar con los valores de las transformaciones de las distribuciones marginales, $\hat{\lambda}_1, \dots, \hat{\lambda}_p$ obtenidas previamente.
- Entonces mediante un proceso iterativo obtenemos un conjunto de valores $\lambda = (\lambda_1, \dots, \lambda_p)$ el cual colectivamente maximiza

$$l(\lambda_1, \dots, \lambda_p) = -\frac{n}{2} \ln |\mathbf{S}(\lambda)| + \sum_{i=1}^p \left[(\lambda_i - 1) \sum_{j=1}^n \ln(x_{ji}) \right]$$

donde $\mathbf{S}(\lambda)$ es la matriz de covarianza muestral

$$\mathbf{S}(\lambda) = \frac{1}{n} \sum_{j=1}^n (\mathbf{x}_j^{(\lambda)} - \bar{\mathbf{x}}^{(\lambda)})(\mathbf{x}_j^{(\lambda)} - \bar{\mathbf{x}}^{(\lambda)})', \quad \bar{\mathbf{x}}^{(\lambda)} = \frac{1}{n} \sum_{j=1}^n \mathbf{x}_j^{(\lambda)}$$

Transformaciones de observaciones multivariadas

- La obtención del vector de parámetros λ , mediante la maximización de $\iota(\lambda_1, \dots, \lambda_p)$ es mas complicada y además no suele aportar mejoras importantes respecto a transformar inividualmente cada variable para que las marginales sean normales.
- Por ejemplo en el caso bivariado, transformar cada distribución marginal a una normal es más o menos equivalente a transformar directamente la distribución bivariada a una distribución normal bivariada.
- En general es más fácil seleccionar apropiadas transformaciones para las distribuciones marginales que para las distribuciones conjuntas.
- **Moraleja:** la transformación de potencias de Box-Cox sobre los vectores de observaciones, es muy complicado y es mejor transformar a normales cada uno de sus componentes