ESTADÍSTICA II
Tema 2: Contrastes no paramétricos

- Contrastes de bondad de ajuste
  - Contrastes basados en la distribución χ^2
  - Contraste de Kolmogorov-Smirnov
  - Contrastes específicos para la distribución normal
  - Gráficos de probabilidad
- Contrastes de homogeneidad e independencia basados en la distribución χ^2

Contrastes no paramétricos

Hipótesis no paramétricas: no se pueden escribir en función de un número finito de parámetros.

1. **Bondad de ajuste:** A partir de una muestra $X_1, ..., X_n \sim F$ de variables aleatorias iid, contrastar:
   - $H_0 : F = F_0$ donde F_0 es una distribución prefijada.
   - $H_0 : F \in \{F_\theta : \theta \in \Theta\}$, donde Θ es el espacio paramétrico.

2. **Homogeneidad:** Dados $X_1, ..., X_n \sim F$ y $Y_1, ..., Y_n \sim G$ de variables aleatorias iid, contrastar $H_0 : F = G$.

3. **Independencia:** Dada $(X_1, Y_1), ..., (X_n, Y_n) \sim F$ de vectores bidimensionales aleatorios iid, contrastar $H_0 : X$ e Y son independientes.

Contraste χ^2 de bondad de ajuste (hipótesis nula simple)

Sea $X_1, ..., X_n$ una muestra de vaiid de $X \sim F$ y F_0 una distribución totalmente especificada. Queremos contrastar $H_0 : F = F_0$.

Ejemplo: Tiramos un dado 100 veces y obtenemos los resultados siguientes:

x_i	1	2	3	4	5	6
Frecuencia	10	20	20	10	15	25

¿Está el dado trucado?

Corresponde a tomar F_0 como la distribución uniforme sobre $\{1, 2, 3, 4, 5, 6\}$.

Contraste χ^2 de bondad de ajuste (hipótesis nula simple):

1. Discretización de H_0: Se define una partición del espacio muestral de X en k clases $A_1, ..., A_k$.
2. Se calculan las frecuencias observadas de datos en cada clase

$$O_j = \#\{i : X_i \in A_j\}, \quad j = 1, \ldots, k.$$

3. Se calculan las frecuencias esperadas si H_0 fuera cierta:

$$e_j = np_j, \quad j = 1, \ldots, k,$$

donde $p_j = \mathbb{P}_{F_0}(A_j)$.

4. El estadístico del contraste (medida de discrepancia entre la muestra observada y la hipótesis nula) es el **estadístico de Pearson** (Pearson 1900),

$$\chi^2 = \sum_{j=1}^{k} \frac{(O_j - e_j)^2}{e_j} = \sum_{j=1}^{k} \frac{O_j^2}{e_j} - n.$$

5. Dado un nivel de significación α, la región crítica del contraste es $R = \{\chi^2 > c_\alpha\}$, donde c_α es tal que $\alpha = \mathbb{P}_{F_0}\{\chi^2 > c_\alpha\}$.

Observaciones:

El resultado del contraste depende de la selección de k y de la partición $A_1, ..., A_k$ (ver Sec. 27.4 y 27.5 de DasGupta 2008). Normalmente se recomienda que $e_j \geq 5$ y $O_j \geq 5$, $j = 1, \ldots, k$. Se suelen escoger valores de k entre 5 y 15.

Distribución de χ^2 bajo H_0

Teorema: Bajo $H_0 : F = F_0$, se verifica

$$\chi^2 = \sum_{j=1}^{k} \frac{(O_j - e_j)^2}{e_j} \xrightarrow[n\to\infty]{d} \chi^2_{k-1}.$$

Observación: Hay una restricción entre los términos $(O_j - e_j)/\sqrt{e_j}$ debido a que $\sum_{j=1}^{k} O_j = n$. Y esto reduce de k a $k-1$ los g.l. de la χ^2.

Demostración:

- Para $i = 1, \ldots, n$ se define el siguiente vector aleatorio en $\mathbb{R}^k$:

$$\begin{aligned} \boldsymbol{\xi}_i &= (0, \overset{j-1}{\ldots}, 0, \overset{j}{1}, 0, \overset{k-j}{\ldots}, 0)', \qquad \text{si } X_i \in A_j. \\ &\sim \text{Multinomial}(1, p_1, \ldots, p_k) \end{aligned}$$

- Se cumple $\mathbb{E}(\boldsymbol{\xi}_i) = \mathbf{p}$ y $\boldsymbol{\Sigma} := \mathbb{V}(\boldsymbol{\xi}_i) = \mathbf{P} - \mathbf{p}\mathbf{p}'$, donde $\mathbf{p} = (p_1, \ldots, p_k)'$ y $\mathbf{P} = \text{diag}(p_1, \ldots, p_k)$.
- Relación entre $\boldsymbol{\xi}_i$ y estadístico de Pearson:

$$\begin{aligned} \frac{1}{\sqrt{n}}\mathbf{P}^{-1/2}\left(\sum_{i=1}^{n}\boldsymbol{\xi}_i - n\,\mathbf{p}\right) &= \left(\frac{(O_1 - e_1)}{\sqrt{e_1}}, \ldots, \frac{(O_k - e_k)}{\sqrt{e_k}}\right)' \\ &= \mathbf{P}^{-1/2}\sqrt{n}(\bar{\boldsymbol{\xi}} - \mathbf{p}). \end{aligned}$$

- Por lo tanto
$$\chi^2 = \|\mathbf{P}^{-1/2}\sqrt{n}(\bar{\boldsymbol{\xi}} - \mathbf{p})\|^2$$

- Por TCL, $\sqrt{n}(\bar{\boldsymbol{\xi}} - \mathbf{p}) \xrightarrow[n\to\infty]{d} \mathrm{N}_k(\mathbf{0}, \boldsymbol{\Sigma})$ y por el teorema de la aplicación continua, $\chi^2 \xrightarrow[n\to\infty]{d} \|\mathbf{Y}\|^2$, donde $\mathbf{Y} \sim \mathrm{N}_k(\mathbf{0}, \mathbf{V})$ y $\mathbf{V} := \mathbf{P}^{-1/2}\boldsymbol{\Sigma}\mathbf{P}^{-1/2}$.

- La matriz $\mathbf{V} = \mathbf{I}_k - \sqrt{\mathbf{p}}\sqrt{\mathbf{p}}'$, con $\sqrt{\mathbf{p}} := (\sqrt{p_1}, \ldots, \sqrt{p_k})'$ es simétrica e idempotente (ejercicio).

- Por lo tanto $\|\mathbf{Y}\|^2 \sim \chi^2_{\mathrm{tr}(\mathbf{V})}$, pero $\mathrm{tr}(\mathbf{V}) = k - 1$.

□

Por el resultado anterior, la región crítica (asintótica) del contraste para un nivel de significación α es:

$$R = \{\chi > \chi^2_{k-1;\alpha}\}.$$

Ejemplo: Tiramos un dado 100 veces y obtenemos los resultados siguientes:

x_i	1	2	3	4	5	6
Frecuencia	10	20	20	10	15	25

Contrasta la hipótesis nula de que el dado no está trucado a nivel $\alpha = 0.05$. ¿Cuál es el p-valor del contraste?

Contraste χ^2 de bondad de ajuste (H_0 compuesta)

Sean $X_1, ..., X_n \sim F$ vaiid. Queremos contrastar

$$H_0 : F \in \{F_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^r\}. \tag{1}$$

1. Se definen las clases $A_1, ..., A_k$ y se calculan las frecuencias observadas $O_1, ..., O_k$.
2. Estimamos $\boldsymbol{\theta}$ mediante $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(X_1, \ldots, X_n)$, el estimador de máxima verosimilitud (e.m.v.).
3. Se calculan las frecuencias esperadas basadas en el valor del e.m.v., es decir, $\hat{e}_j = n\hat{p}_j$ donde $\hat{p}_j = \mathbb{P}_{\hat{\boldsymbol{\theta}}}(A_j)$.
4. El estadístico de Pearson es

$$\chi^2 = \sum_{j=1}^{k} \frac{(O_i - \hat{e}_j)^2}{\hat{e}_i}$$

5. Fisher (1924): ¿Qué distribución tiene este estadístico bajo H_0?

Al estimar r parámetros se introducen r nuevas restricciones sobre el vector

$$\left(\frac{O_1 - e_1}{\sqrt{e_1}}, \ldots, \frac{O_k - e_k}{\sqrt{e_k}} \right).$$

Bajo condiciones de regularidad (ver, por ej., Teorema 7.133, p. 463, de Schervish 1995)

$$\sum_{j=1}^{k} \frac{(O_j - \hat{e}_j)^2}{\hat{e}_j} \xrightarrow[n\to\infty]{d} \chi^2_{k-r-1}.$$

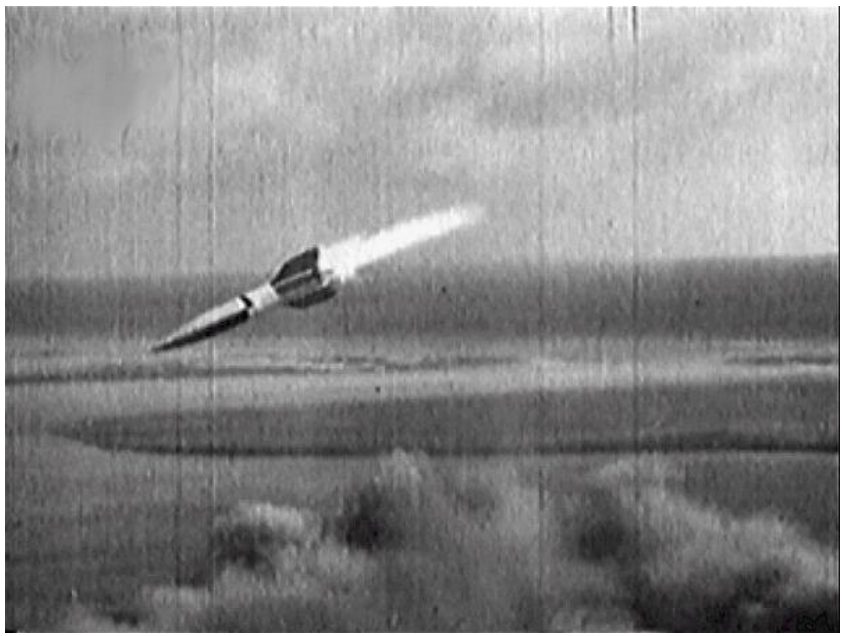
Condiciones:

- Θ es un conjunto abierto de un subespacio afín r-dimensional, con $r < k-1$.
- $\hat{\boldsymbol{\theta}}$ es el e.m.v. de $\boldsymbol{\theta}$ basado en la muestra discretizada $O_1, \ldots, O_k$.
- Si $\boldsymbol{\theta}_0$ es el verdadero valor de $\boldsymbol{\theta}$ y $\hat{\boldsymbol{\theta}}$ es el e.m.v. de $\boldsymbol{\theta}$, entonces $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \to_d \mathrm{N}(\mathbf{0}, \mathcal{I}(\boldsymbol{\theta}_0)^{-1})$, donde $\mathcal{I}(\boldsymbol{\theta}_0)$ es la matriz de información de Fisher.
- Las funciones $\boldsymbol{\theta} \to \mathbb{P}_{\boldsymbol{\theta}}(A_j)$ son inyectivas y derivables dos veces respecto a $\boldsymbol{\theta}$.

Región crítica del contraste (1) para un nivel de significación α:

$$R = \{\chi^2 > \chi^2_{k-r-1;\alpha}\}.$$

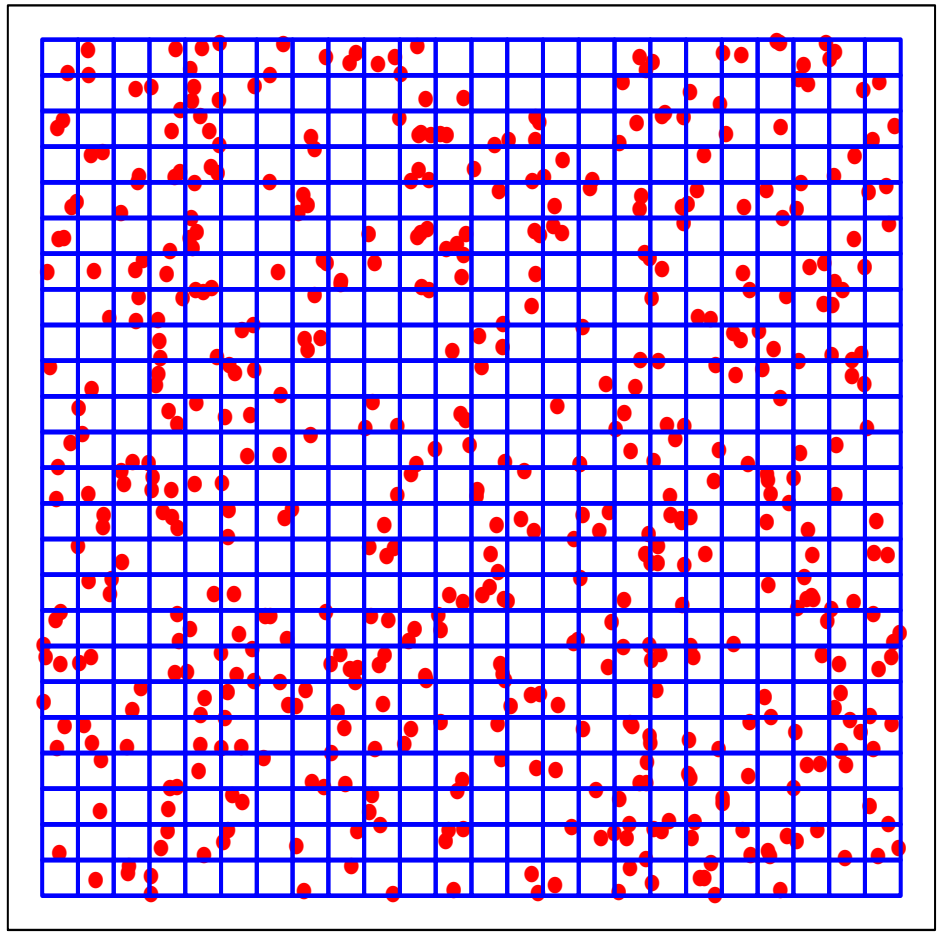
Ejemplo: los bombardeos de Londres



- Desde el 8 de septiembre de 1944 al 27 de marzo de 1945 fueron lanzados contra Inglaterra unos 1.400 misiles V2.

- ¿Eran los lugares de impacto en Londres aleatorios?

- Clarke (1946) dividió un área de 144 km^2 del sur de Londres en 576 cuadrados de 0.25 km^2 cada uno.



- La zona había registrado 537 impactos, por lo que la media era de $537/576 \approx 0.9323$ misiles por cuadrado.

The actual results were as follows:

No. of flying bombs per square	Expected no. of squares (Poisson)	Actual no. of squares
0	226·74	229
1	211·39	211
2	98·54	93
3	30·62	35
4	7·14	7
5 and over	1·57	1
	576·00	576

- Si los lugares de impacto de los V2 eran aleatorios, entonces se podían caracterizar mediante un *proceso de Poisson homogéneo* en un subconjunto de $\mathbb{R}^2$. En ese caso, la distribución de probabilidad del número de bombas en cada uno de los cuadrados

  $N_i =$ "Nº de impactos en el cuadrado i-ésimo", $i = 1, \ldots, 576$,

  debía ser Poisson(λ), con $\lambda > 0$ desconocido.

H_0: La distribución de $N_1, \ldots, N_{576}$ es Poisson de parámetro λ.

Bajo H_0, el estimador de máxima verosimilitud de λ es:

$$\hat{\lambda} = \frac{0 \times 229 + 1 \times 211 + \cdots + 1 \times 7}{576} \approx 0.9323$$

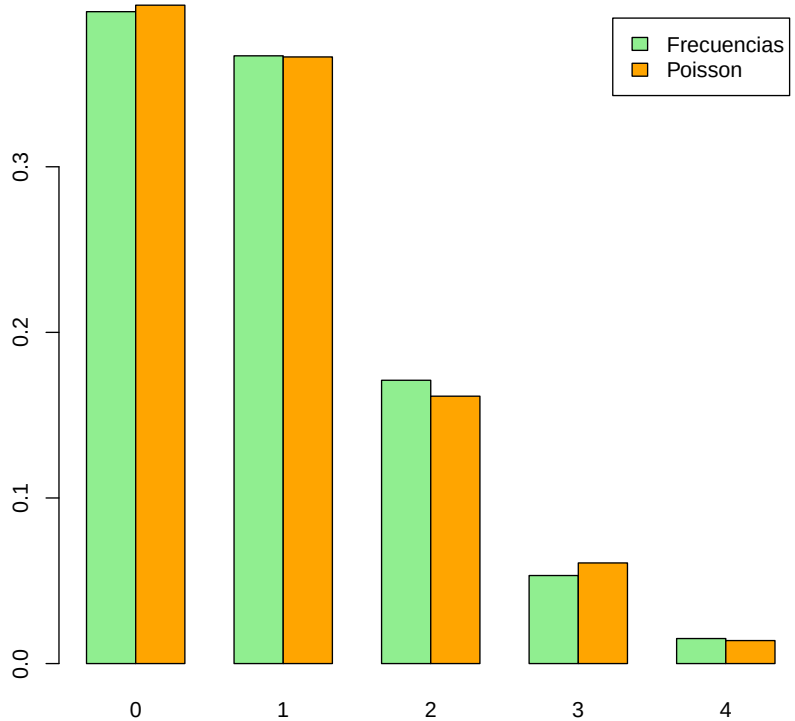
Partición del espacio muestral de $N =$ "Nº de impactos en un cuadrado de 0.25 km^2": $A_1 = \{0\}$, $A_2 = \{1\}$, $A_3 = \{2\}$, $A_4 = \{3\}$, $A_5 = \{4, 5, \ldots\}$.

Frecuencias esperadas estimadas bajo H_0:

$$\mathbb{P}_{\hat{\lambda}}\{N = j\} = e^{-\hat{\lambda}} \frac{\hat{\lambda}^j}{j!}, \qquad j = 0, 1, \ldots$$

$$\hat{e}_j = 576\, \hat{p}_j \quad \text{con } \hat{p}_j = \begin{cases} e^{-\hat{\lambda}} \frac{\hat{\lambda}^j}{j!} & \text{si } 1 \leq j \leq 4, \\ 1 - e^{-\hat{\lambda}} \sum_{j=0}^{3} \frac{\hat{\lambda}^j}{j!} & \text{si } j = 5. \end{cases}$$

j	O_j	$\hat{p}_j$	$\hat{e}_j$
0	229	0.39	226.74
1	211	0.37	211.34
2	93	0.17	98.54
3	35	0.05	30.62
≥ 4	8	0.02	8.71



Estadístico χ^2 de Pearson:

$$\chi^2 = \sum_{j=1}^{5} \frac{(O_j - \hat{e}_j)^2}{\hat{e}_j} = 1.0176$$

Bajo H_0, $\chi^2 \overset{\text{aprox}}{\sim} \chi^2_3$ (5 clases - 1 parámetro estimado -1).

Si $Y \sim \chi^2_3$, entonces $\mathbb{P}\{Y > 1.0176\} \approx 0.797$.

El p-valor del contraste es aproximadamente 0.797.

Al nivel habitual $\alpha = 0.05$ no se puede rechazar que los datos procedan de una distribución de Poisson.

Test χ^2 de bondad de ajuste con R

El comando de R que implementa el test χ^2 es

`chisq.test(datos,p=...)`

- `datos`: Es la muestra de que disponemos
- `p`: Es el vector de probabilidades esperadas.
- Por defecto, se contrasta la hipótesis de que los datos siguen una distribución uniforme.
- Se supone que bajo H_0 la distribución está completamente especificada ($k-1$ grados de libertad)

Entre otras cosas la función calcula el valor del estadístico y el p-valor del contraste.

Para obtener el estadístico χ^2 de Pearson:

```
chisq.test(datos,p=...)$statistic
```

Código R para el ejemplo de los bombardeos:

```
res = c(seq(0,4),7)
obs = c(229,211,93,35,7,1)
n = sum(obs)
lambda = sum(res*obs)/n
masa = dpois(res,lambda)
# Se agrupan las dos ultimas clases:
obs2 = c(obs[1:4],sum(obs[5:6]))
prob = c(masa[1:4],1-sum(masa[1:4]))
esp = n*prob
# Codigo para el grafico de barras:
matriz = rbind(prob,obs2/n)
rownames(matriz) = c("Frecuencias","Poisson")
barplot(matriz,beside=TRUE,names.arg=c(0:4),
        legend.text=TRUE,
        col=c("lightgreen","orange"))
# Test chi 2
t = chisq.test(obs2,p=prob)$statistic
pvalor = 1 - pchisq(t,3)
```

Contraste de Kolmogorov-Smirnov

Sea $X_1, ..., X_n$ una muestra de v.a.i.i.d. de $X \sim F$. Planteamos el contraste

$$H_0 : F = F_0, \tag{2}$$

con F_0 totalmente especificada **y continua**.

Un estimador de $F(x)$ es la función de distribución empírica

$$\mathbb{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}}$$

¿A qué distribución de probabilidad corresponde la función de distribución empírica asociada a la muestra $x_1 = 2$, $x_2 = 4$ y $x_3 = 6$?

Para x fijo, ¿qué distribución tiene la v.a. $n\mathbb{F}_n(x)$? ¿A dónde converge $\mathbb{F}_n(x)$?

Teorema (Glivenko-Cantelli):

$$D_n(F) = \|\mathbb{F}_n - F\|_\infty = \sup_{x\in\mathbb{R}} |\mathbb{F}_n(x) - F(x)| \xrightarrow[n\to\infty]{c.s.} 0.$$

La región de rechazo de Kolmogorov-Smirnov para el contraste (2) es

$$R = \{D_n(F_0) > c_\alpha\},$$

para un valor crítico c_α apropiado.

Lema: Si una v.a. X tiene distribución continua F, entonces $F(X)$ tiene distribución uniforme en $(0, 1)$.

Teorema: Sea $X_1, \ldots, X_n$ una muestra de v.a.i.i.d de X, v.a. continua con función de distribución F. Entonces la distribución de probabilidad de $D_n = \|\mathbb{F}_n - F\|_\infty$ no depende de F.

Demostración:

- Por las propiedades de F como función de distribución, con probabilidad 1,

$$D_n = \sup_{x\in\mathbb{R}} \left| \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{F(X_i)\leq F(x)\}} - F(x) \right| = \sup_{u\in[0,1]} \left| \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{U_i\leq u\}} - u \right|,$$

  donde $U_i = F(X_i) \sim \text{Unif}[0,1]$.

- O también observemos que $D_n = \max\{D_n^+, D_n^-\}$, siendo

$$D_n^+ := \sup_{x\in\mathbb{R}} [\mathbb{F}_n(x) - F(x)] = \max_{1\leq i\leq n} \left[\frac{i}{n} - F(X_{(i)}) \right]$$

  y

$$D_n^- := \sup_{x\in\mathbb{R}} [F(x) - \mathbb{F}_n(x)] = \max_{1\leq i\leq n} \left[F(X_{(i)}) - \frac{i-1}{n} \right].$$

□

En consecuencia, el valor de c_α en la región de rechazo $R = \{D_n > c_\alpha\}$ es el mismo para cualquier distribución F_0 continua.

La distribución de D_n es fácil de simular y además está tabulada (su expresión exacta está, por ej., en la p. 400 de Shao 1999).

Tabla 8 Contraste de Kolmogorov-Smirnov

Valores críticos de $D = |F_n(x) - F(x)|$ donde $F_n(x)$ es la distribución muestral de tamaño n y $F(x)$ la distribución teórica.

Tamaño muestral n	Nivel de significación				
	0,20	0,15	0,10	0,05	0,01
1	0,900	0,925	0,950	0,975	0,995
2	0,684	0,726	0,776	0,842	0,929
3	0,565	0,597	0,642	0,708	0,828
4	0,494	0,525	0,564	0,624	0,733
5	0,446	0,474	0,510	0,565	0,669
6	0,410	0,436	0,470	0,521	0,618
7	0,381	0,405	0,438	0,486	0,577
8	0,358	0,381	0,411	0,457	0,543
9	0,339	0,360	0,388	0,432	0,514
10	0,322	0,342	0,368	0,410	0,490
11	0,307	0,326	0,352	0,391	0,468
12	0,295	0,313	0,338	0,375	0,450
13	0,284	0,302	0,325	0,361	0,433
14	0,274	0,292	0,314	0,349	0,418
15	0,266	0,283	0,304	0,338	0,404
16	0,258	0,274	0,295	0,328	0,392
17	0,250	0,266	0,286	0,318	0,381
18	0,244	0,259	0,278	0,309	0,371
19	0,237	0,252	0,272	0,301	0,363
20	0,231	0,246	0,264	0,294	0,356
25	0,21	0,22	0,24	0,27	0,32
30	0,19	0,20	0,22	0,24	0,29
35	0,18	0,19	0,21	0,23	0,27
>35	$\frac{1,07}{\sqrt{n}}$	$\frac{1,14}{\sqrt{n}}$	$\frac{1,22}{\sqrt{n}}$	$\frac{1,36}{\sqrt{n}}$	$\frac{1,63}{\sqrt{n}}$

n es el tamaño de la muestra.

Ejemplo: Contrasta a nivel $\alpha = 0.01$ si la muestra $16, 8, 10, 12, 6$ procede de una exponencial de media $\lambda = 11.5$ (es decir, $F_0(x) = 1 - e^{-\frac{x}{11.5}}$).

$x_{(i)}$	$\frac{i}{n}$	$F_0(x_{(i)})$	$\frac{i}{n} - F_0(x_{(i)})$	$F_0(x_{(i)}) - \frac{i-1}{n}$
6	0.2	0.41	-0.21	0.41
8	0.4	0.5	-0.1	0.3
10	0.6	0.58	0.02	0.18
12	0.8	0.65	0.15	0.05
16	1	0.75	0.25	-0.05

$$D_n^+ = \max_{1 \le i \le 5} \left[\frac{i}{n} - F_0(x_{(i)}) \right] \quad \text{y} \quad D_n^- = \max_{1 \le i \le 5} \left[F_0(x_{(i)}) - \frac{i-1}{n} \right]$$

$$D_n = \max\{D_n^+, D_n^-\} = 0.41 \quad \text{y} \quad D_{5;0.01} = 0.669.$$

No podemos rechazar H_0.

Contraste de Kolmogorov-Smirnov con R

Comando de R para el test de Kolmogorov-Smirnov:

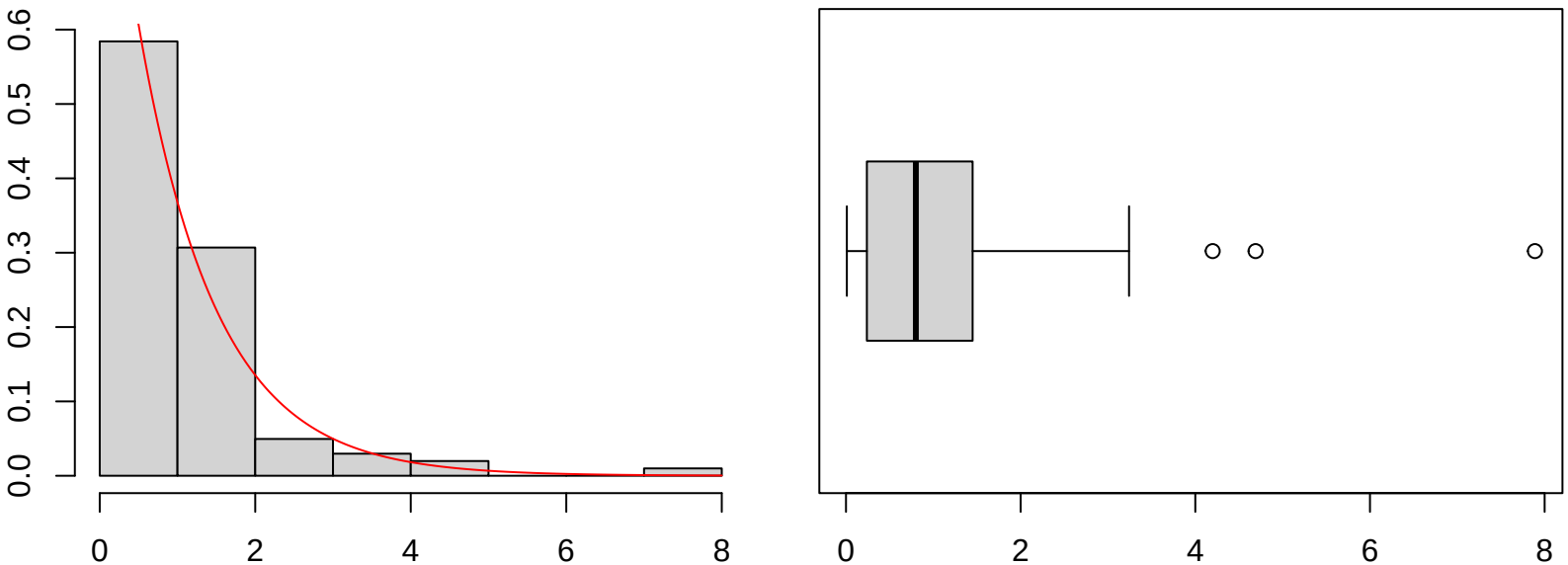
```
ks.test(datos,distribucion,parametros)
```

- `datos`: La muestra de que disponemos
- `distribucion`: Distribución bajo H_0 (por ejemplo, `pnorm`)
- `parametros`: Parámetros de la distribución bajo H_0.

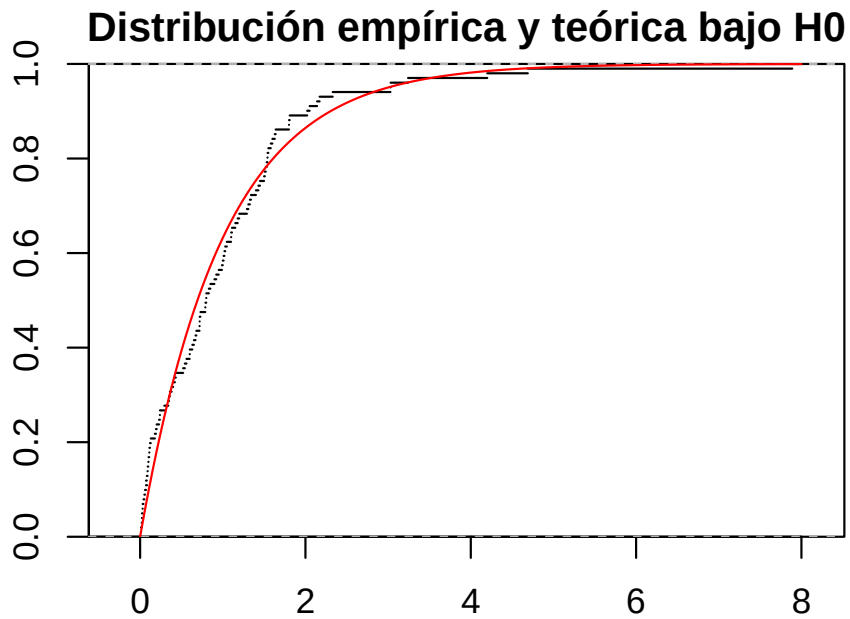
Ejemplo (datos kevlar): Los datos del fichero `kevlar.txt` (procedentes de Barlow *et al.* 1984) corresponden al tiempo hasta el fallo (en horas) de 101 barras de un material utilizado en los transbordadores espaciales, llamado Kevlar49/epoxy, sometidas a un nivel de esfuerzo del 90%.

Vamos a contrastar la hipótesis nula de que los datos tienen distribución exponencial de parámetro $\lambda = 1$.

```
hist(kev, freq=F, main="")
x <- seq(0,8,0.01)
d <- dexp(x)
lines(x,d,col="red")
boxplot(kev,horizontal = T, whisklty = 1)
```



```
Fn <- ecdf(kev)
plot(Fn, do.points = FALSE)
F0 <- pexp(x,1)
lines(x,F0,col="red")
```




```
ks.test(kev,pexp)
        One-sample Kolmogorov-Smirnov test
data:  kev
D = 0.087, p-value = 0.4286
alternative hypothesis: two.sided
```

La desventaja del test de Kolmogorov-Smirnov es que debemos especificar completamente la distribución bajo H_0 (H_0 simple).

La distribución del estadístico $\hat{D}_n = \|\mathbb{F}_n - F_{\hat{\boldsymbol{\theta}}}\|_\infty$, donde $\boldsymbol{\theta}$ denota los parámetros desconocidos del modelo establecido en una H_0 compuesta como la de (1), ya no es conocida ni única para todas las distribuciones continuas y típicamente se aproxima por métodos Monte Carlo.

Otros contrastes de bondad de ajuste son el de Cramér-von-Mises, en el que el estadístico del contraste es

$$C_n(F) = \int [\mathbb{F}_n(x) - F(x)]^2 dF(x),$$

o el de Anderson-Darling

$$A_n(F) = \int [\mathbb{F}_n(x) - F(x)]^2 \frac{1}{F(x)(1-F(x))} dF(x).$$

Los contrastes de bondad de ajuste están relacionados con el problema de *selección de modelos*, en el que el objetivo es seleccionar de entre una colección de modelos probabilísticos, aquél que proporcione el mejor ajuste a los datos (ver Burnham y Anderson 2002).

Contrastes de normalidad

Existen diversos procedimientos (Shapiro-Wilks, K^2 de D'Agostino, etc.) para contrastar si una v.a. X sigue una distribución normal:

$$H_0 : X \sim N(\mu, \sigma^2), \quad \text{para algún } \mu \in \mathbb{R} \text{ y } \sigma > 0. \tag{3}$$

Por simplicidad, aquí describimos el método de Jarque-Bera, que define el estadístico del contraste

$$\text{JB} = \frac{n}{6}\left(S^2 + \frac{1}{4}(K-3)^2\right),$$

donde

$$S = \frac{\hat{\mu}_3}{\hat{\sigma}^3} = \frac{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2\right)^{3/2}} \quad \text{y} \quad K = \frac{\hat{\mu}_4}{\hat{\sigma}^4} = \frac{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2\right)^2}$$

son respectivamente el coeficiente de asimetría y la curtosis.

Bajo la hipótesis nula de normalidad, el estadístico JB sigue asintóticamente una distribución χ^2_2, por lo que, a nivel de significación α, la región de rechazo de H_0 dada en (3) es, para muestras grandes,

$$R = \{\text{JB} > \chi^2_{2;\alpha}\}.$$

Ejemplo:

```
library("moments")

x <- rnorm(1000)
jarque.test(x)
```

```
        Jarque-Bera Normality Test

data:  x
JB = 2.2224, p-value = 0.3292
alternative hypothesis: greater
```

Ejercicio: Generar observaciones de una distribución no normal y ver qué p-valor se obtiene.

Contrastes de normalidad multivariante

Sea $\mathbf{X}$ un vector aleatorio con p componentes. Queremos contrastar

$$H_0: \quad \mathbf{X} \text{ sigue una distribución normal.}$$

Podemos evaluar (parcialmente) la normalidad de $\mathbf{X}$

- examinando las distribuciones de las componentes X_j, $j = 1, \ldots, p$, de $\mathbf{X}$, que deberían ser normales univariantes;
- examinando los diagramas de dispersión bivariantes de las parejas de componentes de $\mathbf{X}$, que deberían tener forma elíptica;
- comprobando si las distancias de Mahalanobis $d_i^2 = (\mathbf{x}_i - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})$ siguen una distribución χ_p^2, por ejemplo, mediante un contraste de bondad de ajuste.

Los contrastes de normalidad multivariante más utilizados se basan en una generalización multivariante (Mardia 1970, 1974) de las medidas de asimetría y curtosis (ver Joenssen y Vogel 2014).

Para $\mathbf{x}_1, \ldots, \mathbf{x}_n$ una muestra de $\mathbf{X}$, vector aleatorio de dimensión $p > 1$, el *coeficiente de asimetría multivariante de Mardia* es

$$\beta_1 = \frac{1}{n^2} \sum_{i,j=1}^{n} (d_{ij}^2)^3,$$

siendo $d_{ij}^2 = (\mathbf{x}_i - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_j - \bar{\mathbf{x}})$ y el *coeficiente de curtosis multivariante* es

$$\beta_2 = \frac{1}{n} \sum_{i=1}^{n} (d_{ii}^2)^2.$$

Si $\mathbf{X}$ es normal multivariante entonces, para n grande,

$$\frac{n}{6} \beta_1 \overset{\text{aprox.}}{\sim} \chi^2_{\text{df}}, \quad \text{con df} = \frac{p(p+1)(p+2)}{6},$$

y

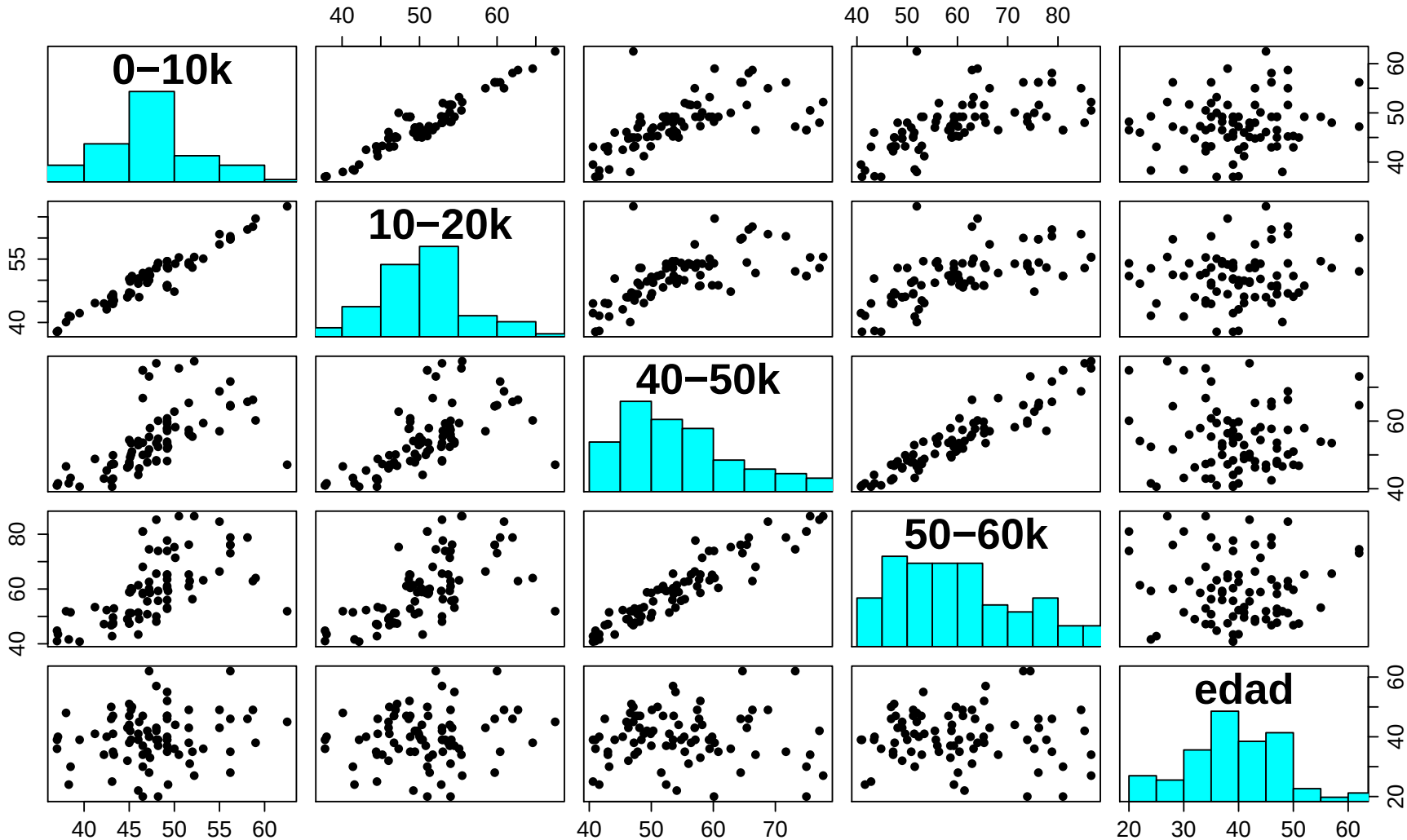
$$\sqrt{n} \, \frac{\beta_2 - p(p+2)}{\sqrt{8p(p+2)}} \overset{\text{aprox.}}{\sim} N(0,1).$$

Ejemplo (ultramaratón): El fichero `race100k.dat` (Everitt 1994) contiene los tiempos (en minutos) que tardan 80 corredores en cubrir cada uno de los 10 tramos de 10 km en una carrera de 100 km. La última columna contiene las edades de los corredores.

```
Race100k = read.table("race100k.dat")
Datos = Race100k[,c(1,2,5,6,11)]

panel.hist = function(x, ...) {
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(usr[1:2], 0, 1.5) )
  h <- hist(x, plot = FALSE)
  breaks <- h$breaks; nB <- length(breaks)
  y <- h$counts; y <- y/max(y)
  rect(breaks[-nB], 0, breaks[-1], y, col = "cyan", ...)
}
pairs(Datos,cex = 1.2, pch = 16,
      diag.panel = panel.hist,
      labels=c("0-10k","10-20k","40-50k","50-60k","edad"),
      cex.labels = 2, font.labels = 2)
```

Ejemplo (ultramaratón):



Ejemplo (ultramaratón):

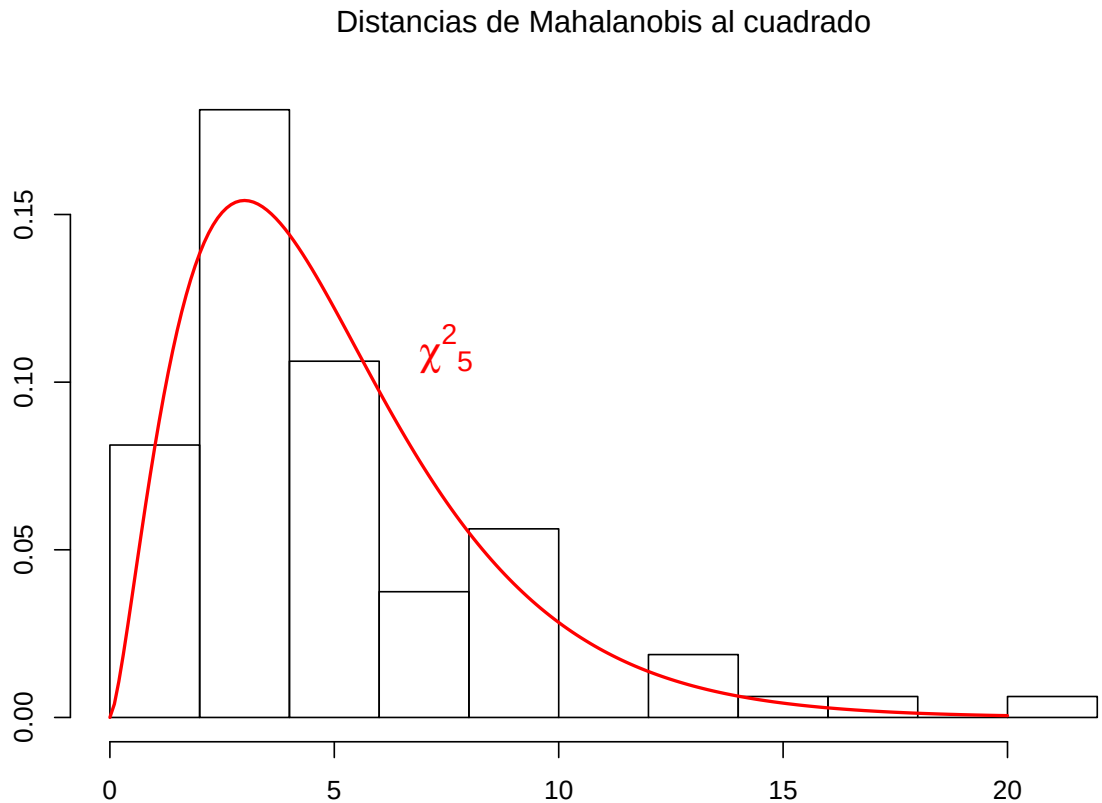
```
# Contrastes de normalidad de Mardia
library("QuantPsyc")
MN = mult.norm(Datos, s = var(Datos))
MN
$mult.test
          Beta-hat     kappa        p-val
Skewness  5.810596 77.474608 4.749059e-05
Kurtosis 38.209231  1.715406 8.627084e-02

$Dsq
 [1]  4.8235127  3.1815515  4.5229703  4.7873549  2.0466464  2.3905665
 [7]  2.1989107  2.4953705  5.7160136  2.8270303  5.7101678  3.7744083
[13]  2.3162065  1.4274329  1.2937829  1.4391935  1.6985519  5.6520187
[19]  2.2567073  3.3072804  6.9196572  2.8106688  1.8698582  1.9781567
[25]  4.6267377  3.4957009  2.7588955  3.2844932  3.2004966  2.8190750
[31]  6.3894767  0.3313730  3.7406764  1.9446374  2.6676929  2.7967326
[37]  1.4025611  4.9620438  2.5514015  4.4742525  3.1004388  8.2481182
[43]  1.8389863  5.7622428  6.3329402  6.4803270  4.3659450  1.7014139
[49]  2.0437456  0.2446679  2.5299637  8.5676925  3.0305199 20.0071382
[55]  4.6611962  5.0467220  1.5908512 12.9457093  4.3002812  8.3244233
[61]  3.3733185  2.9992819 17.5273816  2.5687490 14.3574584  5.9185743
[67]  3.7854423  6.2720720  8.4108593 12.2353951  2.4684825  5.4812627
[73]  9.7492986  8.1520671  4.5078721  9.4548422 12.7913758  9.5383085
[79]  8.8128994  6.5814700
```

Ejemplo (ultramaratón):

```
# Contraste de Kolmogorov-Smirnov sobre distancias de Mahalanobis:
ks.test(MN$Dsq,"pchisq",ncol(Datos))
        One-sample Kolmogorov-Smirnov test

data:  MN$Dsq
D = 0.11764, p-value = 0.2016
alternative hypothesis: two-sided
```



Gráficos de probabilidad

- Es un procedimiento gráfico para ver si una distribución es adecuada para unos datos.
- Si $X \sim F$, entonces se verifica

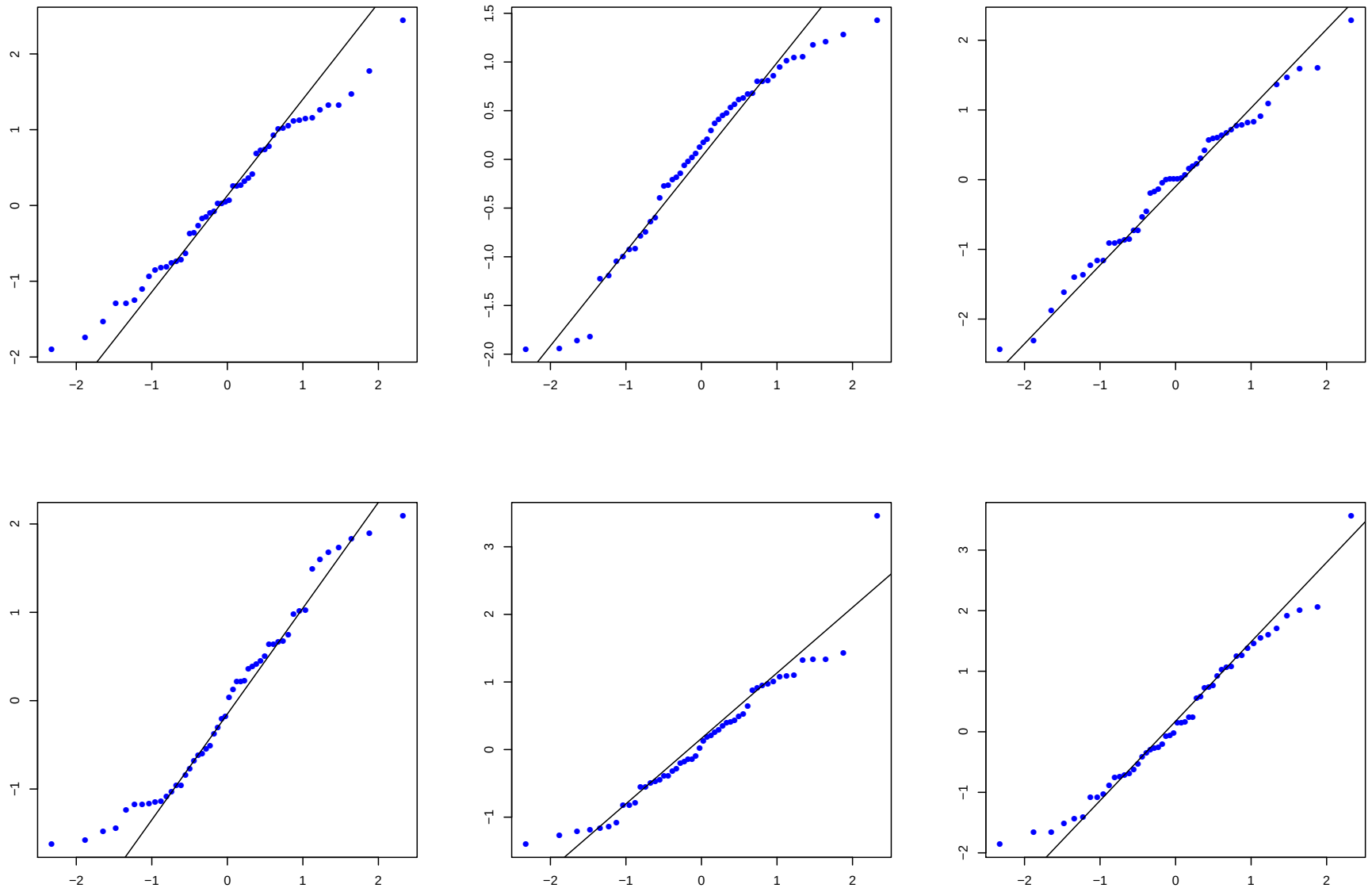
$$\mathbb{E}(F(X_{(i)})) = \frac{i}{n+1}.$$

- Como consecuencia, $X_{(i)} \approx F^{-1}(\frac{i}{n+1})$.
- En particular, para contrastar normalidad, si F y Φ son las funciones de distribución de una $\mathrm{N}(\mu, \sigma^2)$ y una $\mathrm{N}(0,1)$ respectivamente, entonces

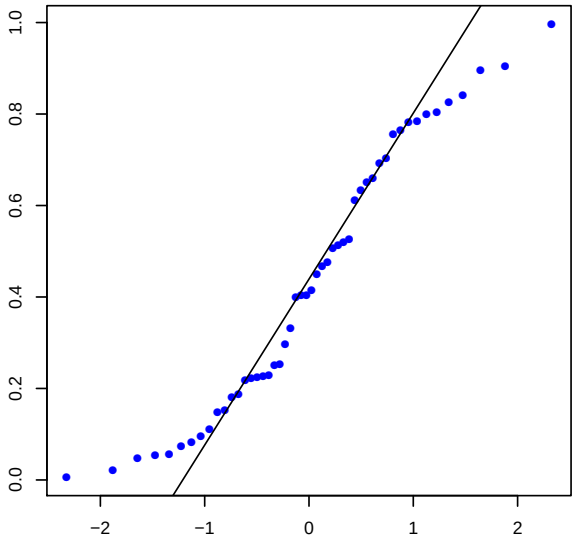
$$X_{(i)} \approx \sigma\, \Phi^{-1}\left(\frac{i}{n+1}\right) + \mu.$$

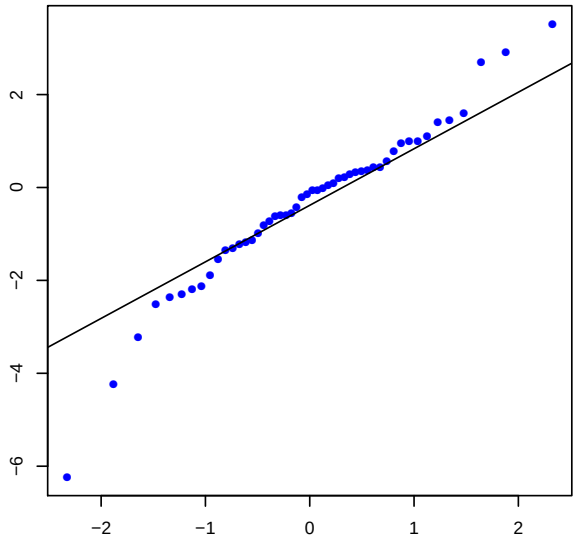
- Se representan los puntos $(X_{(i)}, \Phi^{-1}\left(\frac{i}{n+1}\right))$.

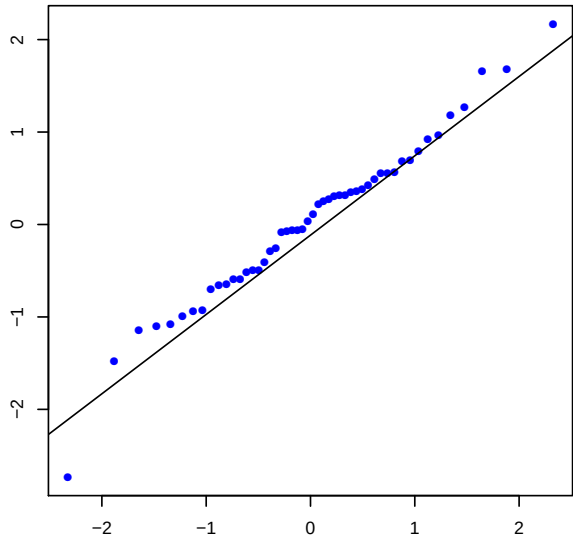
Gráficos de probabilidad para 6 muestras normales ($n = 50$)

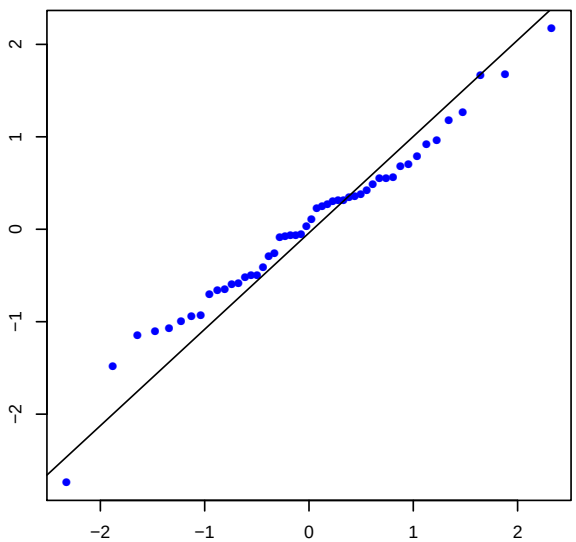


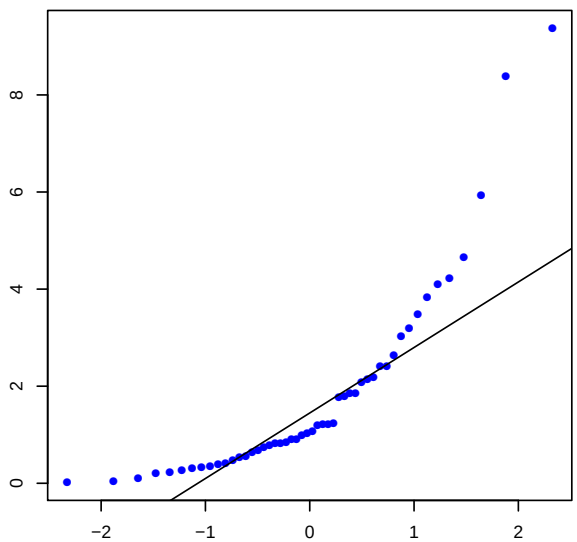
¿Cuáles de estas muestras ($n = 50$) proceden de una distribución normal?

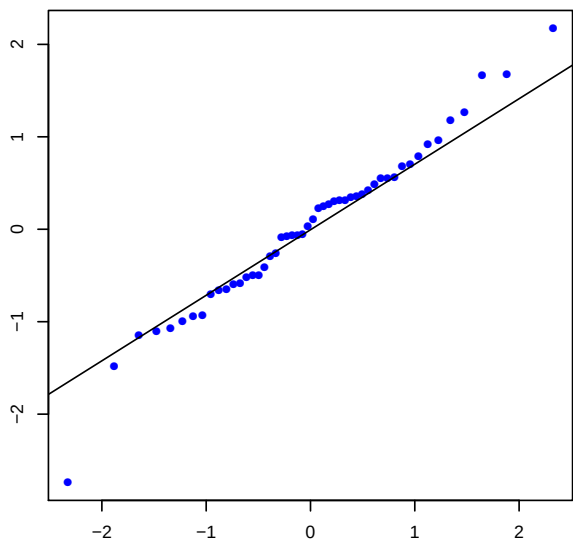






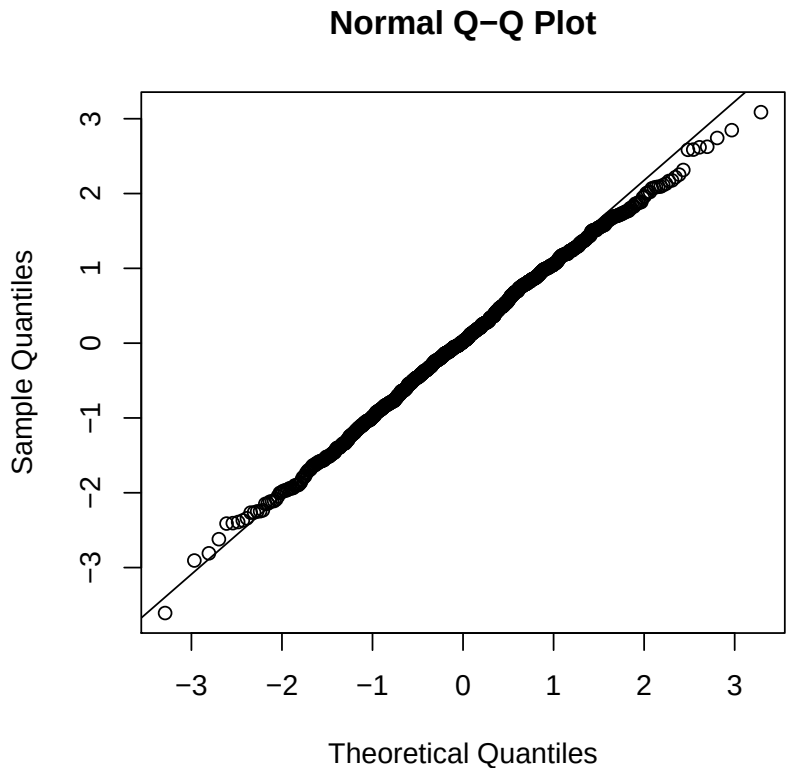






Ejemplo con R:

```
x <- rnorm(1000)
qqnorm(x)
qqline(x)
```



Ejercicio: Generar observaciones de una distribución no normal y ver qué se obtiene. Cambiar los tamaños muestrales y comprobar cómo afecta al gráfico resultante.

Contraste χ^2 de homogeneidad

Para p muestras independientes $M_1, \ldots, M_p$,

$$\left.\begin{array}{rl} M_1 \equiv X_{11}, ..., X_{1n_1} & \overset{iid}{\sim} F_1 \\ \vdots & \vdots \\ M_p \equiv X_{p1}, ..., X_{pn_p} & \overset{iid}{\sim} F_p \end{array}\right\} H_0 : F_1 = F_2 = ... = F_p$$

- Dividimos los datos en clases A_i y consideramos las frecuencias observadas $O_{ij} = \#\{$ datos de M_j en $A_i\}$
- Bajo H_0, $O_{ij} \sim B(n_j, p_i)$, con $p_i = \mathbb{P}_{H_0}(A_i)$.
- Por lo tanto, $e_{ij} = n_j p_i$.

Tabla de contingencia

Las dos matrices siguientes se comparan con el estadístico de Pearson:

	M_1	$\cdots$	M_p
A_1	O_{11}	$\cdots$	O_{1p}
$\vdots$	$\vdots$	$\ddots$	$\vdots$
A_k	O_{k1}	$\cdots$	O_{kp}

	M_1	$\cdots$	M_p
A_1	e_{11}	$\cdots$	e_{1p}
$\vdots$	$\vdots$	$\ddots$	$\vdots$
A_k	e_{k1}	$\cdots$	e_{kp}

- Bajo H_0 se verifica

$$\sum_{j=1}^{p}\sum_{i=1}^{k}\frac{(O_{ij}-e_{ij})^2}{e_{ij}} \xrightarrow[n\to\infty]{d} \chi^2_{p(k-1)}.$$

- Dado que no conocemos las probabilidades p_i, tenemos que estimarlas. Bajo homogeneidad,

$$\hat{p}_i = \frac{\sum_j O_{ij}}{\sum_j n_j} \implies \hat{e}_{ij} = n_j\hat{p}_i = \frac{O_{i\cdot}O_{\cdot j}}{n}.$$

- El estadístico en la práctica (y su dist. bajo H_0) es:

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{p}\frac{(O_{ij}-\hat{e}_{ij})^2}{\hat{e}_{ij}} \xrightarrow[n\to\infty]{d} \chi^2_{(p-1)(k-1)}.$$

- Usar la región crítica

$$R = \{\chi^2 > \chi^2_{(p-1)(k-1),\alpha}\}$$

para hacer el contraste a nivel α.

Ejemplo (Jane Austen): Al morir la escritora Jane Austen (1775–1817) dejó sin concluir su novela *Sanditon*. Desde entonces, varios autores han completado de diferentes formas la novela.

Los datos siguientes (Rice, p. 488) corresponden a la frecuencia de uso de algunas palabras en una muestra de novelas de Austen y en la continuación escrita por un imitador.

Palabra	a	an	this	that	with	without
Imitador	83	29	15	22	43	4
Austen	434	62	86	236	161	38

Sobre la base de estos datos, ¿podemos distinguir entre el estilo de Austen y el estilo del imitador?

Frecuencias esperadas:

Palabra	a	an	this	that	with	without
Imitador	83.54	14.7	16.32	41.69	32.96	6.79
Austen	433.46	76.3	84.68	216.31	171.04	35.21

Diferencias estandarizadas al cuadrado:

Palabra	a	an	this	that	with	without
Imitador	0	13.90	0.11	9.30	3.06	1.14
Austen	0	2.68	0.02	1.79	0.59	0.22

Estadístico de Pearson: $\chi^2 = 0 + 13.90 + \cdots + 0.22 \approx 32.81$

Bajo H_0 χ^2 sigue aproximadamente una distribución χ^2_5.

Mirando las tablas de la χ^2, vemos que $\chi^2_{5,0.05} = 11.07$.
¿Cuál es la conclusión?

Contraste de homogeneidad con R

```
austen=matrix(c(83,434,29,62,15,86,22,236,43,161,4,38),2)

austen
     [,1] [,2] [,3] [,4] [,5] [,6]
[1,]   83   29   15   22   43    4
[2,]  434   62   86  236  161   38

chisq.test(austen)

        Pearson's Chi-squared test

data:  austen
X-squared = 32.8096, df = 5, p-value = 4.106e-06
```

Contraste χ^2 de independencia

Sea $(X_1, Y_1)', ..., (X_n, Y_n)'$ una muestra de vectores i.i.d. de $(X, Y)' \sim F$. Queremos contrastar $H_0 : X, Y$ son v.a. independientes.

Los datos se suelen dar en forma de tabla de contingencia (agrupados):

	B_1	$\cdots$	B_p
A_1	O_{11}	$\cdots$	O_{1p}
$\vdots$	$\vdots$	$\ddots$	$\vdots$
A_k	O_{k1}	$\cdots$	O_{kp}

$$O_{ij} = \text{frecuencia observada de } A_i \times B_j$$

Frecuencias esperadas y su estimación:

$$e_{ij} = np_{ij} = n\,\mathbb{P}\{X \in A_i, Y \in B_j\} \overset{H_0}{=} n\,\mathbb{P}\{X \in A_i\}\,\mathbb{P}\{Y \in B_j\}$$

$$\hat{e}_{ij} = n\frac{O_{i\cdot}}{n}\frac{O_{\cdot j}}{n} = \frac{O_{i\cdot}O_{\cdot j}}{n}.$$

Formalmente, el contraste es exactamente igual al contraste de homogeneidad:

$$R = \{\chi^2 > \chi^2_{(k-1)(p-1);\alpha}\}$$

Ejemplo con R: Los datos son una tabla de contingencia con 13 tareas domésticas y su reparto en la pareja. Las filas son las distintas tareas. Las columnas indican quién realiza la tarea de manera habitual.

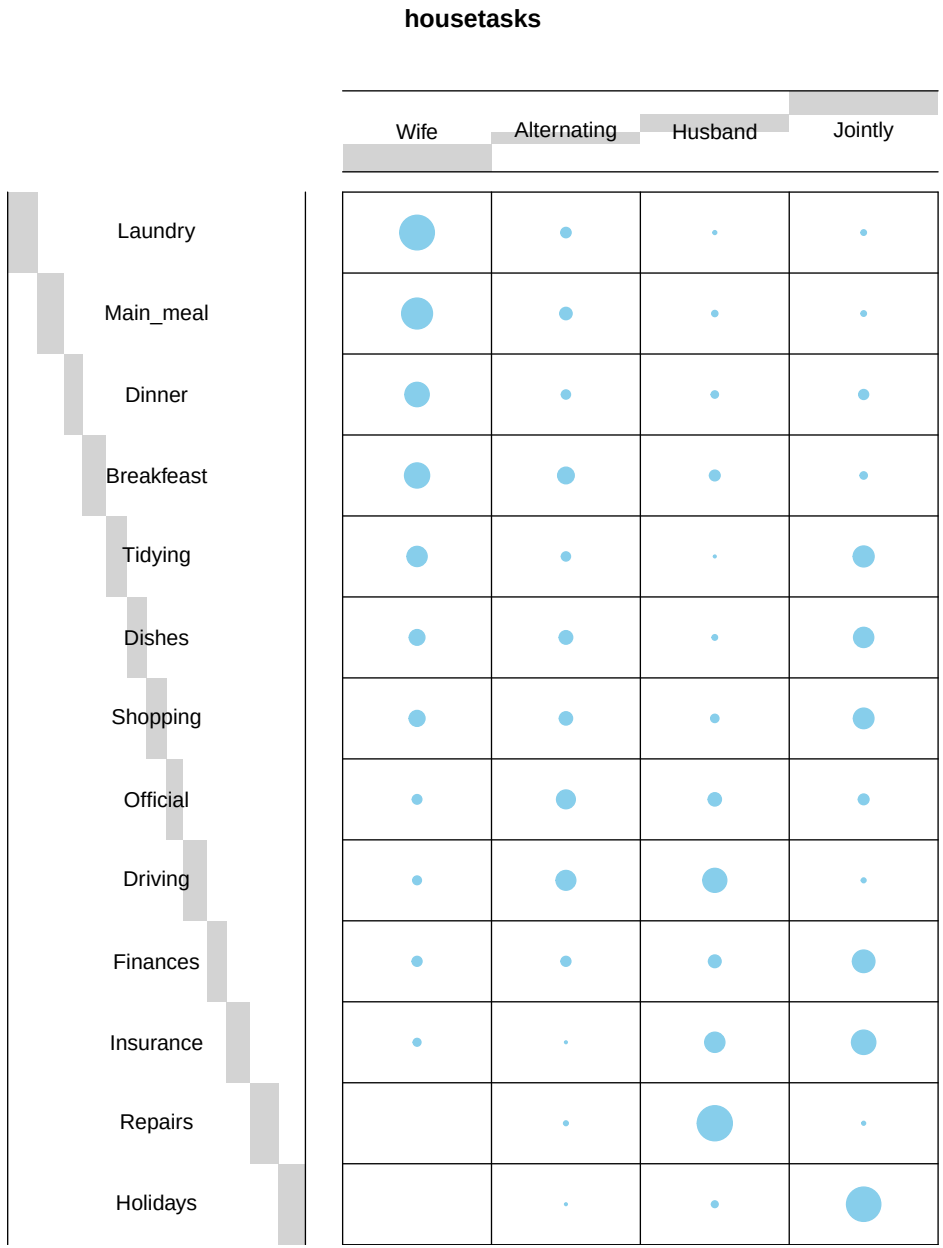
	Wife	Alternating	Husband	Jointly
Laundry	156	14	2	4
Main_meal	124	20	5	4
Dinner	77	11	7	13
Breakfeast	82	36	15	7
Tidying	53	11	1	57
Dishes	32	24	4	53
Shopping	33	23	9	55
Official	12	46	23	15
Driving	10	51	75	3
Finances	13	13	21	66
Insurance	8	1	53	77
Repairs	0	3	160	2
Holidays	0	1	6	153

```
file_path <- "http://www.sthda.com/sthda/RDoc/data/
    housetasks.txt"
housetasks <- read.delim(file_path, row.names = 1)
```

```
chisq <- chisq.test(housetasks)
chisq
        Pearson's Chi-squared test

data:  housetasks
X-squared = 1944.5, df = 36, p-value < 2.2e-16
round(chisq$expected,2) # Frec. esperadas
             Wife Alternating Husband Jointly
Laundry     60.55       25.63   38.45   51.37
Main_meal   52.64       22.28   33.42   44.65
Dinner      37.16       15.73   23.59   31.52
Breakfeast  48.17       20.39   30.58   40.86
Tidying     41.97       17.77   26.65   35.61
Dishes      38.88       16.46   24.69   32.98
Shopping    41.28       17.48   26.22   35.02
Official    33.03       13.98   20.97   28.02
Driving     47.82       20.24   30.37   40.57
Finances    38.88       16.46   24.69   32.98
Insurance   47.82       20.24   30.37   40.57
Repairs     56.77       24.03   36.05   48.16
Holidays    55.05       23.30   34.95   46.70
```

```
library("gplots")
dt <- as.table(as.matrix(housetasks)) # convertimos los datos en tabla
balloonplot(t(dt), main ="housetasks", xlab ="", ylab="", label = FALSE, show.margins = FALSE)
```



Hay otros contrastes no paramétricos (por ejemplo, basados en rangos). Se puede ampliar información en el completo libro de Hollander *et al.* (2013) o en el de Ross (2007).

Referencias

Barlow, R.E., Toland, R.H., Freeman, T. (1984). A Bayesian analysis of stress-rupture life of kevlar 49/epoxy spherical pressure vessels. *Proc. Conference on Applications of Statistics*. Marcel Dekker.

Burnham, K.P., Anderson, D.R. (2002). *Model Selection and Multimodel Inference. A Practical Information-Theoretic Approach*. Springer.

Clarke, R.D. (1946). An application of the Poisson distribution. *Journal of the Institute of Actuaries*, 72, 481.

DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability*. Springer.
Cap. 26: "Goodness of fit"
Cap. 27: "Chi-square tests for goodness of fit"

Fisher, R.A. (1924). The conditions under which χ^2 measures the discrepancey between observation and hypothesis. *Journal of the Royal Statistical Society*, 87, 442–450.

Hollander, M., Wolfe, D.A., Chicken, E. (2013). *Nonparametric Statistical Methods*. 3rd edition. Wiley.

Joenssen, D.W., Vogel, J. (2014). A power study of goodness-of-fit tests for multivariate normality implemented in R. *Journal of Statistical Computation and Simulation*, 84, 1055–1078.

Lehmann, E.L., Romano, J.P. (2005). *Testing Statistical Hypotheses*. Springer.
Cap. 14: "Testing Goodness of Fit"

Mardia, K.V. (1970). Measures of multivariate skewness and kurtosis with applications, *Biometrika*, 57, 519–530.

Mardia, K.V. (1974). Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. *Sankhya: The Indian Journal of Statistics, Series B*, 36, 115–128.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, Series 5, 50, 157–175.

Ross, S.M. (2007). *Introducción a la Estadística*. Ed. Reverté.
Cap. 13: "Contrastes de bondad de ajuste de la chi-cuadrado"
Cap. 14: "Contrastes de hipótesis no paramétricos"

Schervish, J. (1995). *Theory of Statistics*. Springer.
Subsecc. 7.5.2: "Chi-Squared Goodness of Fit Tests"

Shao, J. (1999). *Mathematical Statistics*. Springer.
Subsecc. 6.5.2: "Kolmogorov-Smirnov and Cramér-von-Mises tests"