Handout 12

Goodness of fit: completely specified distribution

1. Table 1 gives the break-out by month of 16766 deaths occurred in the Basque Autonomous Community[1] in 1991. There is a common belief

Table 1: Deaths in the CAPV by month (Year 1991)

Mon	Deaths	Mon	Deaths	Month	Deaths	Mon	Deaths
Jan	1632	Apr	1358	Jul	1231	Oct	1388
Feb	1475	May	1448	Aug	1330	Nov	1380
Mar	1499	Jun	1242	Sep	1256	Dec	1527

that mortality is different on different months (which would make sense if summers of winters are of extreme climatological conditions). The following questions sketch a hypothesis test of that belief.

   (a) What would be the probability of death in each month if the hypothesis: H_0 : "Mortality is uniform all over the year" were true? (Take into consideration that there are months of 28, 30 and 21 days.)

   (b) Under the null hypothesis stated above, what would be the *expected* number of deaths each month?

   (c) Do the data in Table 1 depart enough from the expected monthly deaths to warrant rejection of H_0 at the $\alpha = 0.05$ level? (Hint: you may save yourself work if you think of the distribution of the test statistic under the null hypothesis before doing any computations. You may not need to do all of them to reach a conclusion.)

   (d) Interpret your results. Do you see any reason explaining what you have found?

[1] Data from the Anuario Estadístico Vasco 1992, EUSTAT, Gasteiz/Vitoria, Tabla 2.2/11, pág. 62.

Goodness of fit: non-completely specified distribution

2. Table 2 was compiled by Galton[2]. It classifies 205 couples according to the height of husband and wife.

Table 2: Couples classified according to height of husband and wife

		Wife			
		Tall	Medium	Short	Total
Husband	Tall	18	28	14	60
	Medium	20	51	28	99
	Short	12	25	9	46
	Total	50	104	51	205

(a) Test whether men and women mate irrespective of their height.

(b) Is a test of independence adequate here, or a test of homogeneity? Explain your answer.

References

S. E. Fienberg. *The Analysis of Cross-Classified Categorical Data.* MIT Press, Cambridge, Mass., 1980.

[2]Reproduced in Fienberg (1980), p. 26.

Respuestas abreviadas

1. (a) Las longitudes de los 12 meses de 1991 (no bisiesto) fueron:

```
> d <- c(31,28,31,30,31,30,31,31,30,31,30,31)
```

Podemos comprobar que totalizan 365 días.

```
> sum(d)

[1] 365
```

Las probabilidades de fallecimiento en cada uno de los meses bajo la hipótesis de que son iguales salvo por su longitud son:

```
> p <- d / 365
> sum(p)

[1] 1
```

(b) El número de fallecimientos cada mes y el total fueron:

```
> m <- c(1632,1475,1499,1358,1448,1242,1231,1330,
+        1256,1388,1380,1527)
> sum(m)

[1] 16766
```

El número esperado de fallecimientos cada mes, dado que el total es 16766, es:

```
> e <- p * sum(m)
> e

 [1] 1423.962 1286.159 1423.962 1378.027 1423.962
 [6] 1378.027 1423.962 1423.962 1378.027 1423.962
[11] 1378.027 1423.962
```

(c) Para contrastar la hipótesis indicada, el estadístico Z^2 es:

```
> Z2 <- sum( (m-e)^2 / e )
> Z2

[1] 127.7206
```

que bajo la hipótesis nula se distribuye como χ^2_{12-1}. El valor crítico con $\alpha = 0.05$ es:

```
> qchisq(0.95, df=11)

[1] 19.67514
```

Por tanto, al nivel de significación señalado contundentemente rechazaríamos la hipótesis de que todos los meses son iguales salvo por su longitud.

Si hiciéramos esto manualmente, compondríamos una tabla tal como esta:

```
> tabla.ji <- data.frame(O=m,E=e,diff=m-e,
+                        diff2=(m-e)^2,Z=(m-e)^2/e)
> tabla.ji
      O    E    diff    diff2        Z
1  1632 1424  208.04 43279.96 30.39405
2  1475 1286  188.84 35660.96 27.72671
3  1499 1424   75.04  5630.75  3.95429
4  1358 1378  -20.03   401.10  0.29107
5  1448 1424   24.04   577.84  0.40580
6  1242 1378 -136.03 18503.45 13.42749
7  1231 1424 -192.96 37234.20 26.14831
8  1330 1424  -93.96  8828.79  6.20016
9  1256 1378 -122.03 14890.69 10.80580
10 1388 1424  -35.96  1293.24  0.90820
11 1380 1378    1.97     3.89  0.00282
12 1527 1424  103.04 10616.90  7.45589
```

La suma de los terminos en la última columna proporciona el estadístico de contraste.

```
> sum(tabla.ji["Z"])

[1] 127.7206
```

No nos sería preciso computar más que el primero, porque sólo él (30.39405) ya excede holgadamente de 19.7, el valor crítico delimitando una región crítica de tamaño $\alpha = 0.05$.

(d) Las mayores excesos de fallecimientos sobre los esperables bajo la hipótesis de igual riesgo a lo largo del año se producen en los meses invernales, probable resultado del efecto de infecciones estacionales sobre personas de edad avanzada o salud precaria.

2. Se trata de un contraste de independencia: se han escogido 205 parejas, no cuotas fijas de acuerdo a la talla del marido (lo que supondría fijar el margen derecho) o de acuerdo a la talla de la esposa (lo que supondría fijar los totales en la línea inferior de la Tabla 2).

   Utilizando un procedimiento que reproduce el manual, los cálculos que haríamos serían:

```
> dd <- matrix(c(18,20,12,28,51,25,14,28,9), 3, 3)
> dd

     [,1] [,2] [,3]
[1,]   18   28   14
[2,]   20   51   28
[3,]   12   25    9

> (tfilas <- apply(dd, 1, sum))

[1] 60 99 46

> (tcols  <- apply(dd, 2, sum))

[1]  50 104  51

> (pi.    <- tfilas / sum(dd))

[1] 0.2926829 0.4829268 0.2243902

> (p.j    <- tcols  / sum(dd))

[1] 0.2439024 0.5073171 0.2487805

> (pij    <- pi. %o% p.j)    # producto externo

           [,1]     [,2]       [,3]
[1,] 0.07138608 0.148483 0.07281380
[2,] 0.11778703 0.244997 0.12014277
[3,] 0.05472933 0.113837 0.05582391

> e <- sum(dd) * pij     # esperados en cada casilla
> O <- c(dd)   # "estirar" la matriz en forma de vector
> E <- c(e)    # idem
> tmp <- data.frame(O=O,E=E,diff=(O-E),diff2=(O-E)^2,Z=(O-E)^2/E)
> tmp

   O        E       diff      diff2          Z
1 18 14.63415  3.3658537 11.3289709 0.77414634
2 20 24.14634 -4.1463415 17.1921475 0.71199803
3 12 11.21951  0.7804878  0.6091612 0.05429480
4 28 30.43902 -2.4390244  5.9488400 0.19543465
```

```
5 51 50.22439  0.7756098  0.6015705 0.01197766
6 25 23.33659  1.6634146  2.7669482 0.11856697
7 14 14.92683 -0.9268293  0.8590125 0.05754822
8 28 24.62927  3.3707317 11.3618322 0.46131424
9  9 11.44390 -2.4439024  5.9726591 0.52190755

> Z <- sum(tmp[,5])
> Z

[1] 2.907188

> 1 - pchisq(Z, df=4)    #  p-value

[1] 0.5734754
```

El p-value sugiere que la hipótesis de emparejamiento al azar por lo que respecta a talla no debe ser rechazada.

Orientaciones docentes

1. Se les puede mencionar el uso de `chisq.test`, pero es preferible que al menos las primeras veces hagan los contrastes construyendo el estadístico manualmente —es así como lo habrán de hacer si encuentran una pregunta sobre esto en un examen—.

   El primer ejercicio con ayuda de la función `chisq.test` se reduciría a:

   ```
   > chisq.test(m, p=d/365)

           Chi-squared test for given probabilities

   data:  m
   X-squared = 127.72, df = 11, p-value <
   2.2e-16
   ```

   con los objetos

   ```
   > m

    [1] 1632 1475 1499 1358 1448 1242 1231 1330 1256
   [10] 1388 1380 1527

   > d
   ```

```
 [1] 31 28 31 30 31 30 31 31 30 31 30 31
```

definidos anteriormente.

2. *Una vez que saben hacerlo manualmente*, puede mostrarse el modo de servirse de funciones especializadas de R para hacer el ejercicio 2 con mucha mayor rapidez:

```
> tmp <- loglin(dd, margin=c(1,2))

2 iterations: deviation 2.842171e-14

> 1 - pchisq(tmp$pearson, df=tmp$df)

[1] 0.5734754
```

O bien:

```
> chisq.test(dd)

        Pearson's Chi-squared test

data:  dd
X-squared = 2.9072, df = 4, p-value = 0.5735
```

3. Enfatizar que estamos haciendo contrastes de bondad de ajuste, sin alternativa especificada. Por ejemplo, el rechazo de la nula en el segundo ejercicio no autoriza a decir que "los altos las prefieren altas" (o "las altas, altos"), ni nada por el estilo. Simplemente autorizaría a decir que hombres y mujeres no se emparejan de modo independiente de sus respectivas estaturas.