Handout 5

Case studies to illustrate use of some common distributions

1. Company A is in the business of auto repairs, offering express services. The "star" service is replacement of four tires, which the company claims it is done in time averaging to 65 minutes. When company A applies for a quality assurance certificate, you are selected to perform a quality audit and, among other things, verify their 65-minute mean service time claim.

   You know from experience (and company A is willing to confirm) that service time is exponentially distributed. Unable to check all times for all customers, you collect at random one hundred service times, and find that the total adds up to 6832 minutes. Further, the largest happens to be 312 minutes. With this evidence on hand, you must decide whether the claimed 65-minute mean service time is plausible or not.

   (a) If service time is indeed exponentially distributed, $\exp(\lambda)$, with $m = 65$ minutes, what is λ?

   (b) With said λ, how likely is a service time as large, or larger, as 312 minutes?

   (c) Always with λ as in (a), what would be the probability of no times of 312 minutes, or above, among the one hundred times examined? Do yo think now a 312-minute service time invalidates the company's claim?

   (d) Always with λ as in (a), what is the distribution of the total service time for 100 customers?

   (e) Is the observed total time of 6832 minutes in good agreement with the theoretical distribution in (1d)?

   (f) What would be your decision? Would you accept the company claim, or reject it as implausible given the available evidence?

2. Coastal Insurance Company underwrites insurance for beachfront properties along the Virginia, Carolina and Georgia coasts. It uses the estimate that the probability of a Category III hurricane (sustained winds of more than 110 mph or higher) striking a location of the coast in any one year is 0.05. We will assume the probability of more than one is negligible.

(a) What is the probability that a homeowner purchasing a property with a 30-year mortgage will experience at least one hurricane of Category III during the mortgage period? Use both the binomial and Poisson approximation approach. Would a normal approximation be adequate here?

(b) If Coastal Insurance Company contracts insurance for 1,000 homeowners, in a fairly small area, what is the expected number of claims per year caused by hurricanes of Category III?

(c) What is the *distribution* of the number of claims caused by Category III hurricanes? (Remember all homeowners are "in a fairly small area".)

(adapted from Lind et al. (2011))

Things to investigate on your own.

Case 1 makes some assumptions (and raises some questions) that deserve further thought.

1. Service times are assumed to be exponentially distributed. This is a rather bizarre assumption. Why?

2. Two possible ways of deciding about the company's claim are sketched: one looks only at the largest service time, other at the sum of service times. In both cases, we attempt to assess whether they are plausible given the company's claim.

   This raises some questions: which of the two methods is better? Is there something still better than either? In both cases we are looking at a summary of the sample: the largest and the sum of the service times. Would looking at each service time individually provide some more information?

   You will see systematic treatment of these questions in a few weeks, as we deal with so-called *sufficient statistics*. For now, squeeze your intuition, see what it tells you.

References

D. A. Lind, W. G. Marchal, and S. A. Wathen. *Statistical Techniques in Business and Economics*. McGraw Hill Higher Education, 2011.

Soluciones abreviadas

1. (a) Si $m = 65$ minutos, $\lambda = 1/65$.

   (b)
```
> 1 - pexp(312, rate=1/65)
[1] 0.008229747
```
   Es *muy raro* que el tiempo de servicio alcance los 312 minutos si de verdad la distribución es exponencial y $\lambda = 1/65$. Acontece con probabilidad menor al 1%.

   (c) Sin embargo, las cosas cambian cuando examinamos cien tiempos de servicio. La probabilidad de *algún* tiempo de servicio de 321 minutos o aún mayor es:
```
> 1 - (pexp(312, rate=1/65))^100
[1] 0.5623693
```
   ("1 menos la probabilidad de que *todos* los cien tiempos sean inferiores a 312"). Esta probabilidad no es en absoluto pequeña. La probabilidad de que *un* tiempo alcance ó supere 312 es muy pequeña, pero la probabilidad de que *alguno entre 100 tiempos* alcance o exceda 312 es tan grande como 0.5623693.

   (d) La suma de los 100 tiempos de servicio es una $\gamma(a = \lambda = 1/65, r = 100)$.

   (e) Es tanto como preguntar si 6832 ó mayor es un valor que puede razonablemente presentarse al muestrear una $\gamma(a = 1/65, r = 100)$. La probabilidad de un tal valor es:
```
> 1 - (pgamma(6832, rate=1/65, shape=100))
[1] 0.2961434
```
   ¡Ojo! Es fácil meter la pata. Lo que es a en nuestra notación,

$$f(x) = \frac{a^r}{\Gamma(r)} x^{r-1} e^{-ax},$$

   R lo llama `rate`; lo que es r en nuestra notación, R lo llama `shape`. En un examen no tendrían acceso a R ni a tabletas/teléfonos, y la distribución $\gamma(a, r)$ no les sería accesible. Sin embargo, han de recordar que si el tiempo total de servicio T se distribuye como

$$T \sim \gamma(a = \lambda = 1/65, r = 100),$$

   entonces

$$2\lambda T \sim \gamma(a = 1/2, r = 100)$$

   que es una χ^2_{200}. Entonces la probabilidad que queremos es:

```
> X <- 2 * (1/65) * 6832
> 1 - pchisq(X, df=200)
```

```
[1] 0.2961434
```

resultado idéntico al obtenido con la $\gamma(a, r)$ (de hecho, entre bambalinas R utiliza la distribución gamma para hacer los cálculos correspondientes a la χ^2).

Esta aproximación tampoco les es accesible, puesto que las tablas sólo llegan a $n = 100$. Pero $\chi^2_n \xrightarrow{d} N(n, 2n)$. Utilizando esta aproximación, obtenemos

```
> 1 - pnorm(X, mean=200, sd=sqrt(400))
```

```
[1] 0.3047563
```

Conclusión: 6812 no es un valor en absoluto raro cuando $\lambda = 1/65$. Un valor tal o aún mayor se presentaría con probabilidad cercana al 30% aunque el verdadero λ fuera el indicado.

(f) Ninguno de los dos resultados considerados en (c) y (e) puede considerarse evidencia en contra de la alegación de la compañía de que el tiempo de servicio medio es de 65 minutos. Nótese que el resultado (b) no puede utilizarse.

2. (a) La probabilidad utilizando la distribución binomial es:

```
> 1 - pbinom(0, prob=0.05, size=30)
```

```
[1] 0.7853612
```

Utilizando la aproximación de Poisson, tendríamos $\lambda = 30 \times 0.05 = 1.5$) y la probabilidad buscada sería aproximadamente:

```
> 1 - ppois(0, lambda=1.5)
```

```
[1] 0.7768698
```

Como se ve, una bastante buena aproximación.

(b) El valor medio del número de peticiones de indemnización por año sería $m = np = 1000 \times 0.05 = 50$.

(c) Cabe imaginar que un huracán afectaría a todas o la mayor parte de las viviendas en un área pequeña como la descrita en el enunciado. Por tanto, la distribución *no* seria $b(p = 0.05, n = 1000)$. Seguiría siendo cierto que $m = 50$, pero ningún año se presentarían 50 indemnizaciones: habría (muchos) años de 0 y (pocos) años de 1000.

Pautas docentes

1. En estos y todos los problemas que hagamos es importante ir inculcando gradualmente el modo empirista de hacer del estadístico: dada una conjetura ("¿Es el tiempo medio de servicio de 65 minutos?") transformarla en una conjetura sobre el o los parámetros de un modelo ("¿Es $\lambda < 1/65$?") y ver si la evidencia empírica entra o no en conflicto con dicha conjetura. Esto nos facilitará pronto introducir la lógica del contraste de hipótesis.

2. En el primer problema hay que destacar que podríamos examinar la evidencia de diferentes maneras: tomando el total de tiempos de servicio T, o el tiempo promedio (equivalente), o el mayor de ellos. Es útil dejar planeando sobre sus cabezas la pregunta: ¿Qué sería lo óptimo? ¿Perdemos algo utilizando uno u otro, habría un tercero que es aún mejor? Ello anticipa la cuestión que resolveremos más tarde al introducir la noción de suficiencia.

3. Como en handouts anteriores, es didáctico discutir los supuestos y mostrar cuáles son manifiestamente inadecuados. La independencia entre peticiones de indemnización en el problema 2 es claramente un supuesto inadecuado. La distribución exponencial de los tiempos de servicio, también[1].

[1] Por la falta de "memoria" de la distribución exponencial. La probabilidad de que un operario tarde más de t minutos en cambiar nuestros neumáticos es:

$$P(T \geq t) = 1 - (1 - e^{-\lambda t}) = e^{-\lambda t}.$$

Si llegamos al taller y nos dicen que el operario lleva ya k minutos con nuestros neumáticos y k es grande, sería razonable pensar que ya falta poco. Sin embargo, bajo el modelo exponencial, la probabilidad de que tarde todavía t minutos más es:

$$P(T \geq t + k | T \geq k) = \frac{e^{-\lambda(t+k)}}{e^{-\lambda k}} = e^{-\lambda t}$$

¡La probabilidad de que acabe en t minutos es la misma que si acabara de empezar! ¡De modo bien contraintuitivo, no nos aprovecha nada saber el tiempo que el operario lleva ocupado con nuestro coche!