Muestreo por distancias

El Problema:

Estimar los siguientes parámetros:

- **N:** Tamaño de una población.
- Densidad **D** (Nº de objetos por unidad de área)

Relación entre D y el tamaño de la población **N** en un área **A:**

$$N = D \cdot A$$

¡¡OJO!! Esta relación sólo es válida si los objetos de la población se distribuyen de manera homogénea sobre la región **A.**

Solución 1. Aproximación clásica

1. Seleccionar al azar en la región de estudio (de área **A**), un conjunto de zonas (usualmente circulares, cuadradas o rectangulares) de área total a.
2. Contar el número total n de objetos en el conjunto de zonas seleccionado.
3. Estimar la densidad de objetos mediante:

$$\hat{D} = \frac{n}{a}$$

4. Estimar el tamaño de la población mediante:

$$\hat{N} = \hat{D} \times A$$

Dificultades de esta aproximación:

- Requiere la capacidad de poder contar ***todos*** los objetos dentro de cada una de las zonas seleccionadas. Normalmente en poblaciones biológicas no todos los objetos de una zona pueden ser contados (se esconden, huyen, son demasiado pequeños para verlos todos, …)

- La distribución de los objetos sobre el terreno puede no ser homogénea.

- La selección aleatoria de zonas puede producir que éstas estén alejadas entre sí, o que requieran rutas complejas para poder visitarlas todas, dificultando o encareciendo el proceso de visitarlas.

Solución 2. Aproximación mediante el muestreo por distancias

1. Seleccionar al azar un conjunto de líneas (transectos lineales) o puntos (transectos puntuales) sobre la región en la que se encuentra la población de interés.

2. Medir las distancias desde el transecto a todos los objetos detectados a medida que el observador viaja a lo largo de las líneas, o se sitúa en los puntos elegidos.

3. Estimar la **función de detección g(y)**. Esta función representa la probabilidad de que un objeto situado a distancia y del observador sea detectado por éste. Usualmente esta función es decreciente con y.

4. Si $\hat{g}(y)$ es el estimador de la función de detección, *w* es la distancia máxima de observación, *L* la longitud total del transecto (lineal) recorrido, y *n* el número total de objetos detectados, el estimador de la densidad de objetos es:

$$\hat{D} = \frac{n}{2L\int_0^w \hat{g}(y)\,dy}$$

5. El estimador del tamaño de la población es, nuevamente:

$$\hat{N} = \hat{D} \times A$$

Ventajas del muestreo por distancias: permite estimar el tamaño de la población en aquellos casos en que no todos los objetos pueden ser detectados. Por sorprendente que pueda parecer, es posible obtener buenas estimaciones incluso cuando 80-90% de los objetos permanece sin detectar (aunque estos sólo es verdad si se dan determinadas condiciones)

Observaciones:

- En un transecto lineal de longitud L, si la distancia máxima (perpendicular al transecto) a la que el observador es capaz de detectar objetos es *w*, el área total muestreada será:

$$a = 2 \times w \times L$$

- En un transecto puntual, formado por *k* puntos, si la distancia máxima a la que el observador es capaz de detectar objetos es *w*, el área total muestreada será:

$$a = k \times \pi \times w^2$$

- La función de detección puede depender no sólo de la distancia a que se encuentre el objeto, sino de otras variables que influyan en la capacidad de detección del observador:
  - El tamaño del objeto.
  - La sensibilidad del aparato que, en su caso, se utilice.
  - Las condiciones de visibilidad.
  - …
- Poblaciones de comportamiento gregario: cuando los sujetos de la población forman rebaños, manadas, cardúmenes, etc., el objeto del muestreo por distancias no es el *sujeto individual* sino el *grupo*. Se miden entonces las distancias desde el transecto hasta el centro del grupo. Si de cada grupo se cuenta el número de sujetos que lo componen, se puede estimar el tamaño medio $\hat{N}_g$ de los grupos; mediante el muestreo por distancias se estima la densidad media de grupos $\hat{D}_g$. El producto de ambas cantidades nos permite estimar la densidad media de la población:

$$\hat{D} = \hat{D}_g \times \hat{N}_g$$

Tipos de datos en el muestreo por distancias:

1. Datos de distancia exactos: Se dispone de las distancias perpendiculares x_i exactas desde el objeto al transecto; o bien se dispone de las distancias r_i y ángulos θ_i de avistamiento, a partir de los cuales se calculan las distancias exactas mediante $x_i = r_i \cos(\theta_i)$

2. Datos de distancia agrupados: Muchas veces es difícil calcular con exactitud la distancia, y ésta se computa sólo en intervalos: por ejemplo, de 0 a 20 metros, de 20 a 50, de 50 a 100, y de 100 a 200.

3. Datos con truncamiento: son los que se obtienen cuando los objetos a una distancia mayor que cierto valor w_0 son ignorados.

Hipótesis para la validez de la inferencia estadística a partir de un muestreo por distancias.

1. Los objetos se distribuyen espacialmente sobre la región de interés de acuerdo con algún proceso estocástico con parámetro de densidad D = Número de objetos por unidad de área.

2. Los transectos (puntos o líneas) se colocan al azar sobre la región de interés. De hecho no es necesario que los objetos se distribuyan completamente al azar (de acuerdo con un proceso de Poisson), pero los transectos sí que deben estar situados completamente al azar con respecto a la distribución de los objetos.

   Por ejemplo, si la población estuviese dispuesta a lo largo de unos ejes rectilíneos, y los transectos siguiesen precisamente estos ejes, el estimador obtenido sobreestimaría el tamaño de la población, ya que el transecto no visita las zonas más vacías.

3. Para estimar de modo fiable la función de detección deben darse además las siguientes condiciones:

    a. Los objetos que está justo en el transecto son detectados con probabilidad 1 (se detectan siempre)

    b. Los objetos se detectan en su posición inicial, antes de cualquier movimiento en respuesta al observador.

    c. Las distancias se miden con precisión, tanto cuando se miden datos de distancia exactos como cuando se miden datos de distancia agrupados.

    d. Los objetos se identifican correctamente (i.e. no se confunden sujetos pertenecientes a especies distintas, por ejemplo)

    e. Los objetos están inmóviles, o su velocidad con respecto al observador es lenta. Ello garantiza que los objetos no se cuentan dos o más veces (o al menos que son pocos los que se cuentan varias veces).

El procedimiento de muestreo tiene que diseñarse de modo que se cumplan todas estas hipótesis (exactamente, o del modo más aproximado posible), ya que los métodos de análisis de los datos no pueden corregir el efecto de una muestra mal tomada.

Veremos no obstante como y cuanto pueden aminorarse los efectos del no cumplimiento parcial de estas hipótesis. En la actualidad el diseño de métodos de muestreo por distancia sigue siendo un activo campo de investigación.

Transectos lineales. Ideas básicas:

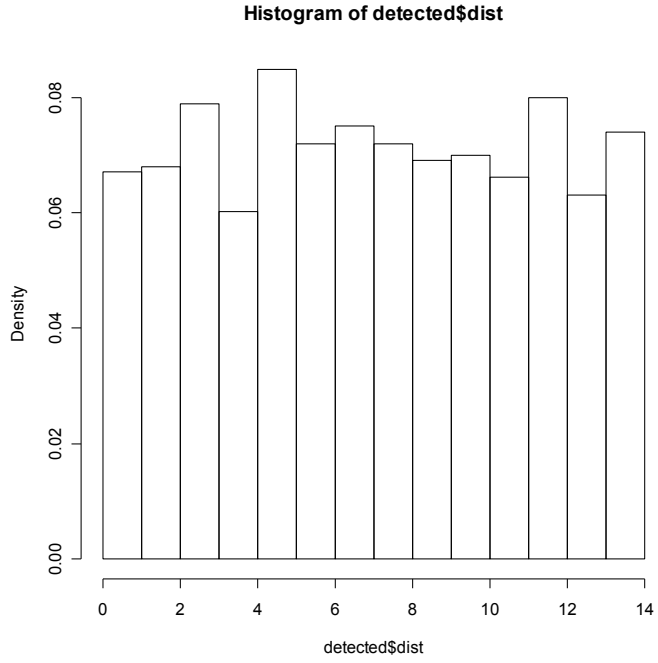
Supongamos:

- que se recorre un conjunto de transectos lineales de longitud total *L.*
- que se detectan **todos** los objetos que se encuentran a una distancia menor o igual que *w* de cada transecto.
- Que en total se han detectado *n* objetos.

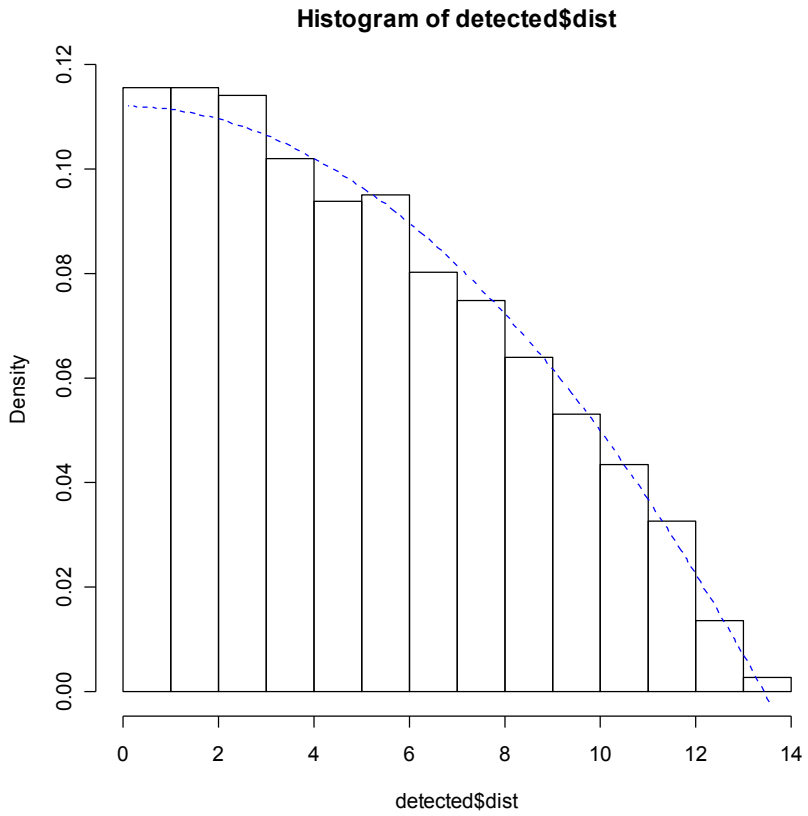
En estas condiciones el área total inspeccionada es a = 2*wL,* por lo que la densidad estimada es simplemente:

$$\hat{D} = \frac{n}{2wL}$$

Si la densidad de objetos es constante en toda la región de interés, el número de objetos detectado a una distancia de entre 0 y 1 u.d. (unidades de distancia: m., km., millas, etc.) debe ser similar al número detectado entre 1 y 2 u.d., y similar al detectado entre 2 y 3 u.d., etc. El histograma de frecuencias para el número de sujetos detectado en cada intervalo de distancia debería ser similar al siguiente:



Supongamos ahora que a medida que los objetos están más lejos es más difícil que sean detectados. En tal caso, el histograma sería más bien como éste:



Dado que el área del histograma representa la frecuencia observada, el cociente entre el área total en el segundo caso (en que se van detectando menos objetos a medida que aumenta la distancia), y el área total en el primer caso (en que se detectan todos los objetos), nos daría una estimación de la proporción de objetos que se dejan de detectar. A continuación precisaremos esta idea.

La línea azul que hemos trazado en la gráfica corresponde a una estimación de la función de densidad $f(x)$ de la variable:

X= “Distancia a que se encuentra un sujeto que ha sido detectado”

Esta función de densidad $f(x)$ debe ser proporcional a la función de detección $g(x)$, esto es $f(x) = c \times g(x)$. En efecto:

$$f(x)\,dx = P(\text{un sujeto esté en } (x, x+dx) / \text{el sujeto es detectado}) =$$

$$= \frac{P(\text{el sujeto es detectado} / \text{el sujeto esté en } (x, x+dx))\, P(\text{el sujeto esté en } (x, x+dx))}{P(\text{el sujeto es detectado})} =$$

$$= \frac{g(x) \times \left(\frac{dx \times L}{w \times L} \right)}{P_a}$$

Por tanto:

$$f(x) = \frac{g(x)}{w \times P_a}$$

De aquí se sigue que:

$$f(0) = \frac{g(0)}{w \times P_a} \Rightarrow w \times P_a = \frac{g(0)}{f(0)}$$

Si conociéramos el valor exacto de P_a, llamando δ al número de sujetos detectados y *n* al número total de sujetos en la zona de área *a*, se tendría:

$$\delta = n \times P_a \quad \Rightarrow n = \frac{\delta}{P_a}$$

Por tanto, la densidad en la zona sería:

$$D = \frac{n}{2wL} = \frac{\delta}{2wLP_a} = \frac{\delta \times f(0)}{2 \times L \times g(0)}$$

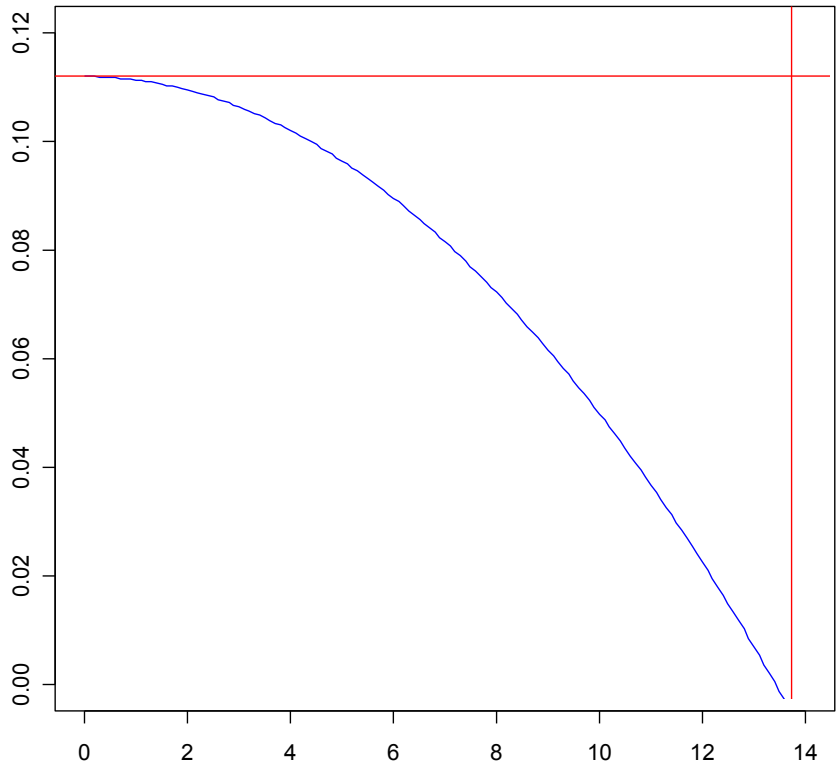
y podríamos estimar:

$$\hat{D} = \frac{\delta \hat{f}(0)}{2L\hat{g}(0)}$$

Observemos por último que, en el caso particular de que $g(0) = 1$:

$$P_a = \frac{1}{w \times f(0)}$$

coincide precisamente con el cociente de dividir el área total bajo la función de densidad $f(x)$ (área bajo la curva azul en la gráfica), que vale 1, entre el área del cuadrado, que vale precisamente $w \times f(0)$, lo que coincide con el razonamiento empírico que hemos hecho más arriba:



Transectos puntuales.

Si el muestreo se realiza mediante transectos puntuales, y consideramos:

- que el observador se sitúa sucesivamente en *k* puntos (transectos).
- que se detectan **todos** los objetos que se encuentran a una distancia menor o igual que *w* de cada punto.
- que en total se han detectado *n* objetos.

Ahora el área total inspeccionada es $a = k\pi w^2$, por lo que la densidad estimada sería:

$$\hat{D} = \frac{n}{k\pi w^2}$$

Si sólo se detecta una proporción P_a de los objetos situados en esa área, del mismo modo que antes, si δ es el número de objetos efectivamente detectados:

$$\hat{D} = \frac{\delta}{k\pi w^2 P_a}$$

Veamos ahora como calcular P_a. Si se detectasen todos los objetos, el número de objetos situados en un anillo a distancia entre *r* y *r+Δ* del observador sería:

$$v = D\pi\left(r+\Delta\right)^2 - D\pi r^2 = D\pi\left(\Delta^2 + 2r\Delta\right)$$

Ahora bien, como la proporción de objetos que se detectan a distancia *r* es $g(r)$, en ese anillo el número de objetos que efectivamente se detectan sería:

$$v_{\text{det}} = D\pi\left(\Delta^2 + 2r\Delta\right) g\left(r\right)$$

Podemos calcular entonces la densidad de puntos detectados a distancia *r* del observador como:

$$f\left(r\right) = \lim_{\Delta\to 0}\frac{v_{\text{det}}}{\Delta} = \lim_{\Delta\to 0} D\pi\left(\Delta + 2r\right) g\left(r\right) = 2D\pi r g\left(r\right)$$

Para calcular ahora el número total de objetos detectados en un círculo de radio *w*, habrá que sumar los detectados en todos los anillos de anchura infinitesimal *dr* desde 0 hasta *w*. Esto es equivalente a calcular la integral:

$$N_{\text{det}}\left(w\right) = \int_0^w 2D\pi r g\left(r\right) dr = 2\pi D\int_0^w r g\left(r\right) dr$$

A su vez, el número de objetos en ese círculo será:

$$N(w) = D\pi w^2$$

Por tanto, la proporción de objetos detectados en un círculo de área $a = \pi w^2$ es:

$$P_a = \frac{N_{\text{det}}(w)}{N(w)} = \frac{2\pi D\int_0^w rg(r)\,dr}{D\pi w^2} = \frac{2}{w^2}\int_0^w rg(r)\,dr$$

Sustituyendo en la expresión de la densidad:

$$\hat{D} = \frac{\delta}{k\pi w^2 P_a} = \frac{\delta}{2k\pi\int_0^w rg(r)\,dr}$$

y llamando:

$$\eta = \pi\int_0^w rg(r)\,dr$$

tendríamos:

$$\hat{D} = \frac{\delta}{2k\eta}$$

Como η no se conoce, deberá estimarse:

$$\hat{D} = \frac{\delta}{2k\hat{\eta}} = \frac{\delta}{2k\pi\int_0^w r\hat{g}(r)\,dr}$$

Estimación de la función de detección $g(x)$

Una de las estrategias habituales para estimar $g(x)$ consiste en parametrizar esta función y estimar sus parámetros. Cualquiera que sea el modelo que se use para parametrizar $g(x)$, debe tener las siguientes características:

- Flexibilidad: debe ser posible ajustar el modelo a la variedad de formas que puede adoptar $g(x)$. Ello excluye en general modelos uniparamétricos.

- Robustez frente a la combinación de estimaciones: es deseable que si hay factores que afecten a la probabilidad de detección, éstos no compliquen en exceso la estimación de $g(x)$.

- Forma de "hombro": consideraciones empíricas indican que la función de detección no debe variar mucho en las proximidades del transecto (i.e., la probabilidad de detección se mantiene alta, próxima a 1, cerca del transecto lineal o puntual). En términos matemáticos ello significa que debería ser $g'(0) = 0$.

- Eficiencia: a la hora de elegir entre dos posibles funciones para modelar la probabilidad de detección será preferible aquella que dé lugar a estimaciones con menor varianza (más precisas).

Una vez que se estima la función de detección, se puede usar el test de la Ji-cuadrado para decidir si los datos se ajustan o no a la función. Si las distancias de detección se dividen en *u* intervalos y llamamos n_i al número de objetos detectados dentro del intervalo *i*-ésimo, y $\hat{E}(n_i)$ al número de objetos que se esperan detectar en ese intervalo de acuerdo con la función elegida, la variable:

$$\chi^2 = \sum_{i=1}^{u} \frac{\left(n_i - \hat{E}(n_i)\right)^2}{\hat{E}(n_i)}$$

sigue una distribución χ^2 con *u*-*m*-1 grados de libertad, siendo *m* el número de parámetros estimados de $g(x)$.

Nota: el test de la χ^2 no permite usualmente elegir entre modelos alternativos, ya que puede darse el caso de modelos que ajusten bien a un conjunto de datos según este test, y sin embargo los modelos sean distintos.

Algunos modelos usuales para la función de detección $g(x)$

Los modelos que se presentan a continuación son de la forma:

$$g(x) = A(x)\left[1 + B(x)\right]$$

$A(x)$	$B(x)$
$\frac{1}{w}$	$\sum_{j=1}^{m} a_j \cos\left(\frac{j\pi x}{w}\right)$
$\frac{1}{w}$	$\sum_{j=1}^{m} a_j \left(\frac{x}{w}\right)^{2j}$
$\exp\left(-x^2/2\sigma^2\right)$	$\sum_{j=2}^{m} a_j \cos\left(\frac{j\pi x}{w}\right)$
$\exp\left(-x^2/2\sigma^2\right)$	$\sum_{j=2}^{m} a_j H_{2j}\left(\frac{x}{\sigma}\right)$ (H = polinomio de Hermite)
$1-\exp\left(-(x/\sigma)^{-b}\right)$	$\sum_{j=2}^{m} a_j \cos\left(\frac{j\pi x}{w}\right)$
$1-\exp\left(-(x/\sigma)^{-b}\right)$	$\sum_{j=2}^{m} a_j \left(\frac{x}{w}\right)^{2j}$