# Principios de Estadística

Leonardo Collado Torres y María Gutiérrez Arcelus
Licenciatura en Ciencias Genómicas, UNAM
www.lcg.unam.mx/~lcollado/index.php
www.lcg.unam.mx/~mgutierr/index.php

Cuernavaca, México
Febrero - Junio, 2009

# Letras y universo

# Intro



- Lo que vamos a ver hoy es un problema curioso relacionado a conteo y tamaño de muestras.
- Pero primero vamos a ver un par de cosas de `R`.

# Gráficas



- Una parte muy importante de `R` es poder visualizar tus datos con diferentes tipos de gráficas. Para esto existen muchos tipos de funciones que se dividen en:
  - *bajo nivel* porque son funciones que pueden graficar encima de gráficas previas. Por ejemplo, lines.
  - *alto nivel* porque siempre crean un nuevo espacio gráfico. Por ejemplo, hist.
- Para ver un índice de las funciones básicas escriban:

```
library(help="graphics")
```

# par



- Otro mundo de diversidad es el de los parámetros de estas funciones. En sí muchos están definidos por par.
- Chequen la ayuda de esta función.
  1. ¿Cúal es el parámetro para ponerle título a una gráfica?
  2. ¿Qué parámetro usarían para definir los límites del eje $Y$?

# plot



- La función de gráficas que vamos a usar hoy es plot. Con esta pueden graficar puntos fácilmente.
- Primero chequen su ayuda y luego definan $x$ y $y$.

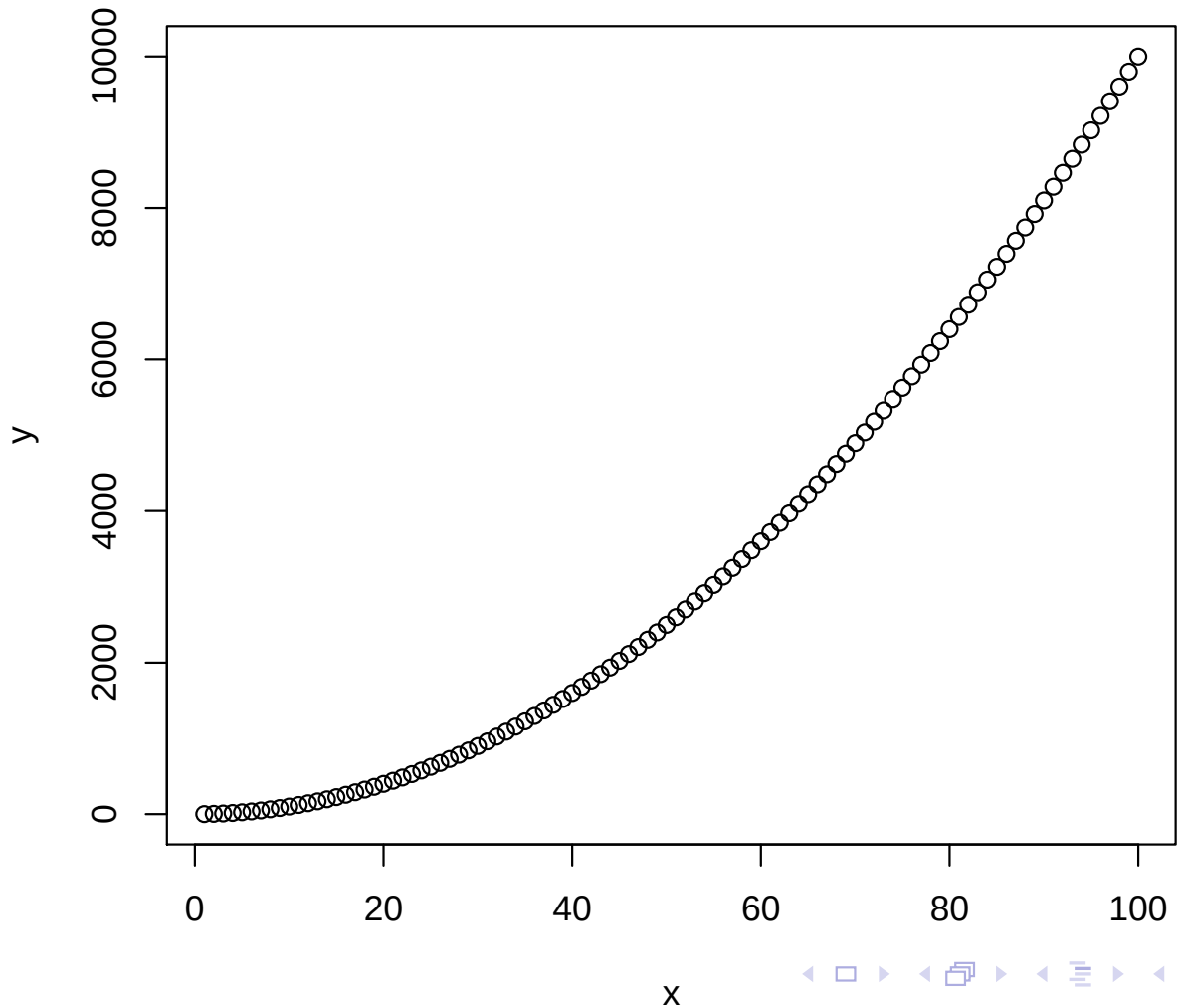
```
> x <- 1:100
> y <- (1:100)^2
```

- Ahora grafiquen los puntos.

```
> plot(x, y)
```

# plot(x,y)





# Plot mejorada



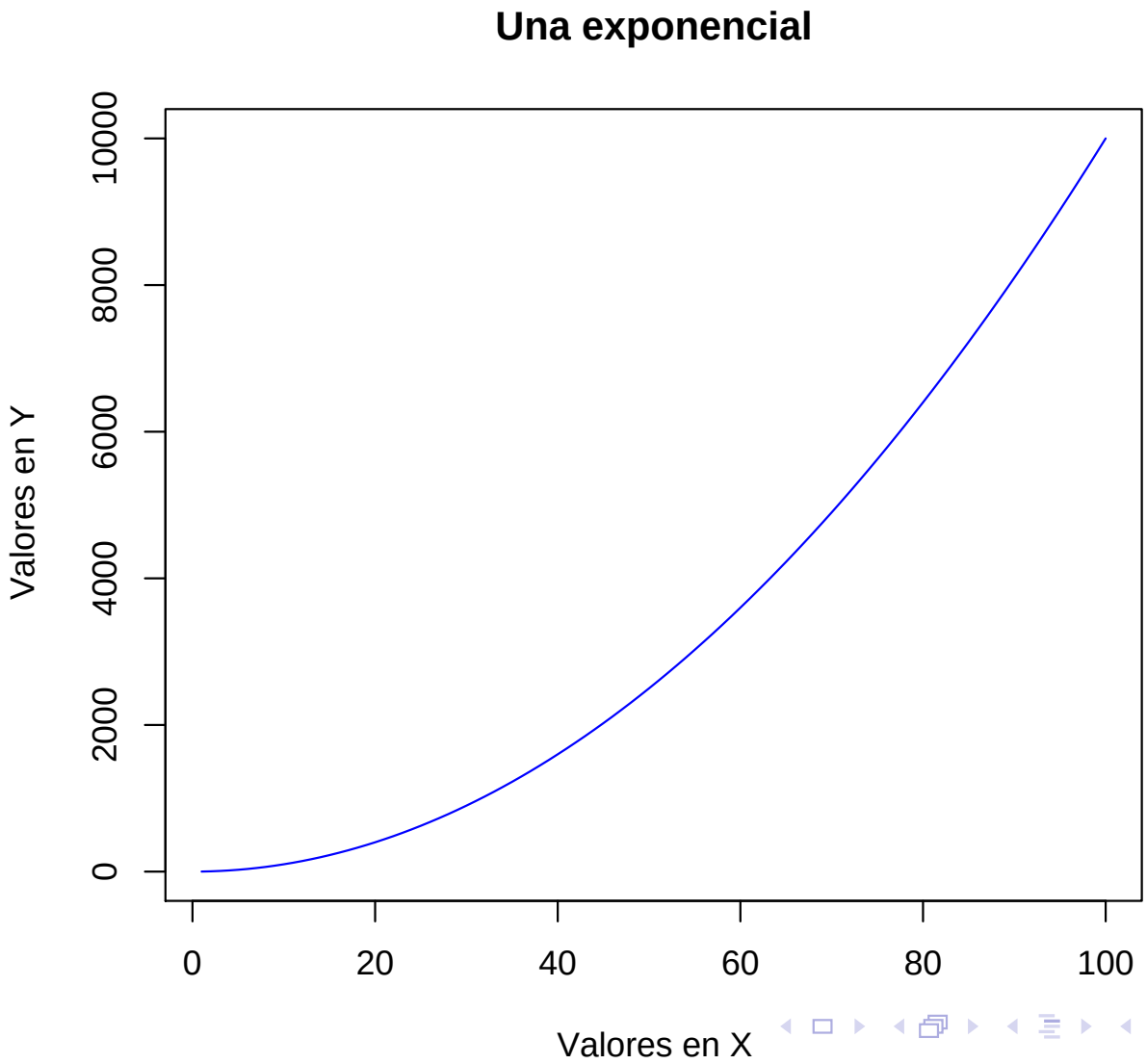
- Ahora hacemos una gráfica más completa

```
> plot(x, y, main = "Una exponencial",
+      ylab = "Valores en Y", xlab = "Valores en X",
+      col = "blue", type = "l")
```

- ¿Qué hace el argumento `type="l"`?

# Plot mejorada





# Pequeña revisión del for



- Solo para que se acuerden :)

```
> res <- NULL
> for (i in 1:10) {
+     if (i == 1) {
+         res <- c(res, runif(1,
+             0, 10))
+     }
+     else {
+         res <- c(res, res[i - 1]^(1/i))
+     }
+ }
```

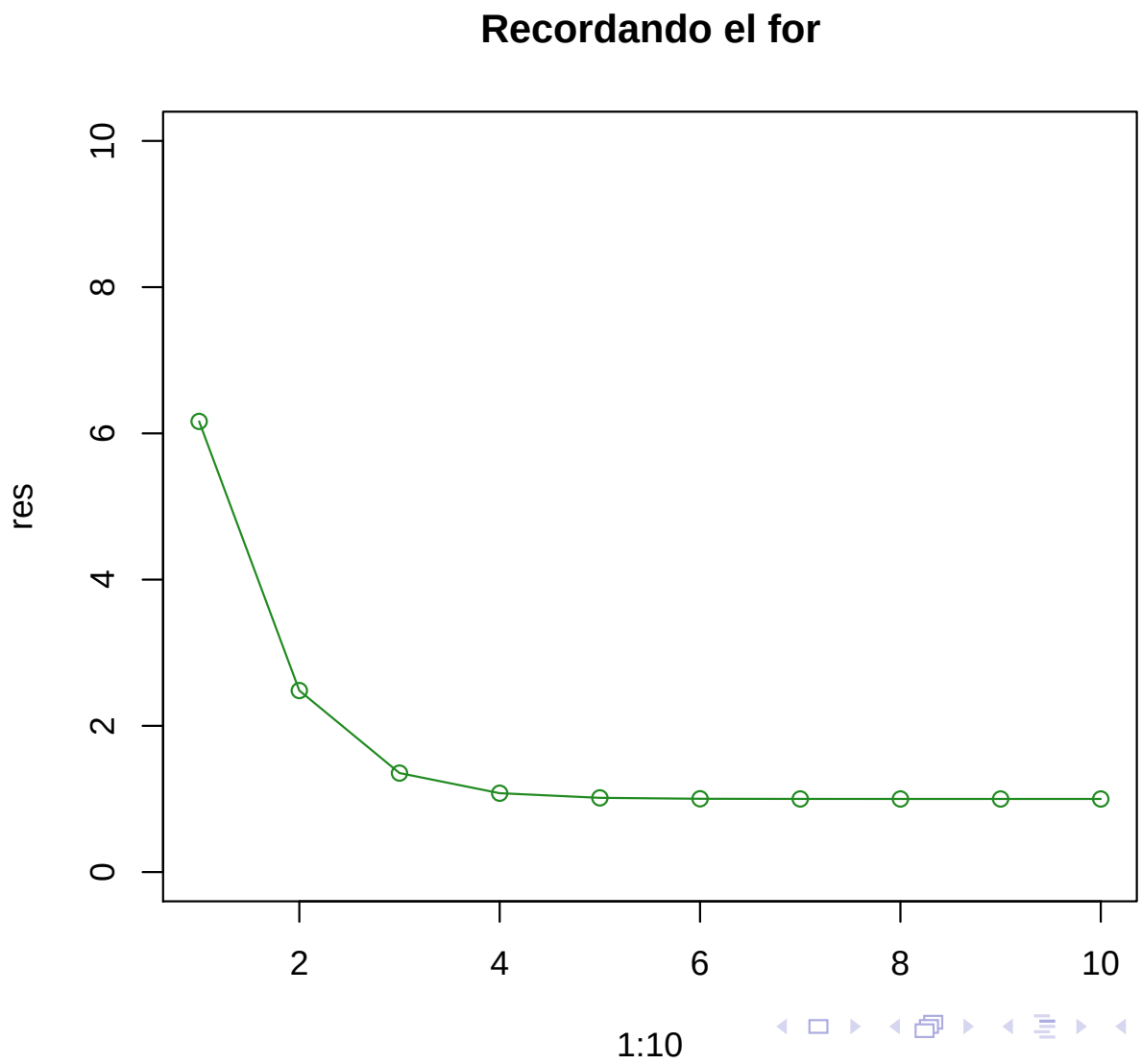
# Pequeña revisión del for



```
> plot(1:10, res, main = "Recordando el for",
+      type = "o", ylim = c(0, 10),
+      col = "forestgreen")
```

# Recordando el for





# El origen



- Bueno, ya con la super intro podemos ahora plantear y resolver el problema.
- Todo surgió porque Osam está buscando cosas raras que pasen con los números al azar. Lo que me planteó recientemente es que si tienes un alfabeto de posibilidades (o números) de tamaño 50...
  - Si sacas 5 elementos al azar, no esperarías que ninguno se repita.
  - Si sacas 50, no esperarías tener uno de cada uno, pues es probable que se repita al menos uno.
  - Si sacas 1000, esperarías tener todos tus elementos al menos una vez con varios repetidos.
- Lo que queremos saber es que tan grande tiene que ser tu muestra para que tengas todos tus elementos al menos una vez. ¿Alguien sabe?

# Caso específico



- Luego de hablar con Osam, nos pusimos Sur y yo a intentar encontrar la respuesta.
- Digamos que tienes $k$ número de elementos (en el caso anterior era 50) y $h$ es el número de elementos que sacamos al azar.
- Digamos que $k = 2$ y $h = 2$. ¿Cúal es la prob. de que aparezcan al menos una vez tus dos elementos? Pues con esta $k$ y $h$ hay 4 casos en donde en 2 se cumple lo que quieres. Osea tu probabilidad es de 2/4 ó 0.5

# Una fórmula



- ¿Qué pasa cuando $k = 2$ y $h = 3$? Tienes 8 casos en los cuales se cumple lo que buscas en 6. Solo hay 2 casos donde o todos son águila o todos son sol (si fuera una monéda), así que tu prob. es de $6/8$. Con $k = 2$ y $h = 4$ tu prob es de $14/16$.
- La probabilidad de que con una muestra de tamaño $h$ aparezcan al menos una vez tus $k$ elementos es igual a uno menos la probabilidad de que no aparezcan.
- Generalizando, la fórmula que te da tu probabilidad es:

$$Prob = 1 - \frac{\sum_1^k (1 - P(k_i))^h}{k - 1}$$

# Una fórmula



- Donde $P(k_i)$ es la probabilidad de que aparezca el elemento $k_i$. Por ahora digamos que todo elemento tiene la misma probabilidad, que es $1/k$.

# A trabajar :)



- Quiero que hagan un barrido de parámetros de la siguiente forma.
- Examinen a las $k$ desde 2 hasta 100.
- Para cada $k$ examinen las $h$ desde 1 hasta 1000.
- Para cada $k$, ¿cuál es la $h$ a partir de donde nuestra probabilidad[1] es de 0.95 o mayor?
- Para alguna $k$, grafiquen las probabilidades en el eje $Y$ y las $h$ en el eje $X$.
- Grafiquen su resultado con las $k$ en el eje $Y$ y las $h$ determinantes en el eje $X$.
- ¿Qué es lo que notan?

[1] De que aparezcan al menos una vez los $k$ elementos.

# Unos tips



- Una forma de resolverlo es con:
  - 2 ciclos tipo `for`.
  - 2 objetos para almacenar sus resultados. Acuérdense de definirlos como `NULL` antes.
  - Acuérdense de las funciones which y sum.

# Respuesta



- Así lo pueden resolver:

```
> res.k <- NULL
> for (k in 2:100) {
+     res.h <- NULL
+     for (h in 1:1000) {
+         no.salir <- NULL
+         for (i in 1:k) {
+             no.salir <- c(no.salir,
+                 (1 - 1/k)^h)
+         }
+         res.h <- c(res.h, 1 - (sum(no.salir)/(k -
+             1)))
+     }
+     res.k <- c(res.k, which(res.h >=
```

# Respuesta



```
+           0.95)[1])
+ }
> head(res.k)

[1]  6  9 12 15 18 21
```

# Gráficas

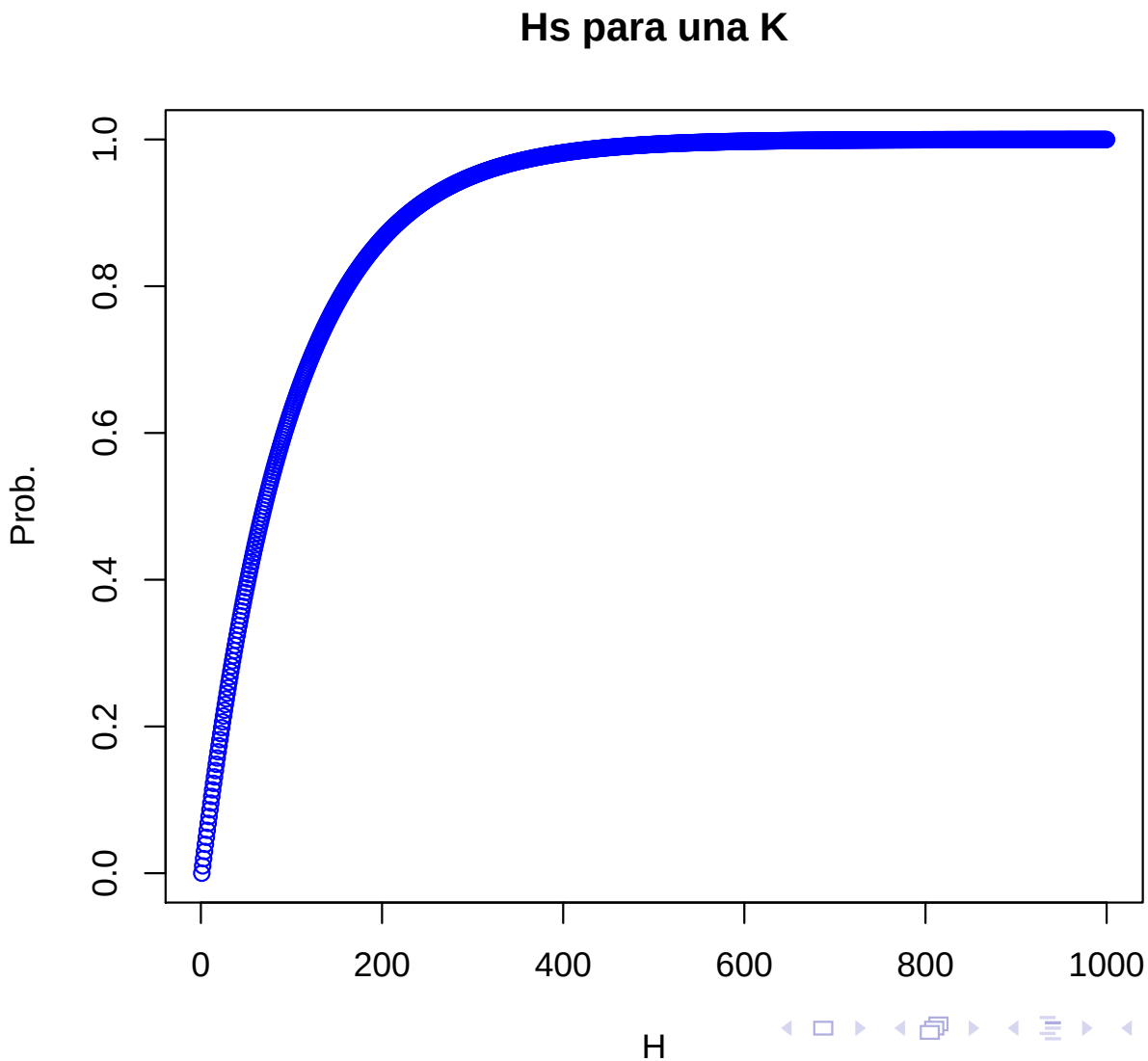


- Ya solo viendo el `head(res.k)` pueden darse cuenta de hacia donde vamos...

```
> plot(res.h, lty = 2, xlab = "H",
+      ylab = "Prob.", main = "Hs para una K",
+      col = "blue")
> plot(res.k, 2:100, xlab = "H determinante",
+      ylab = "K", main = "H determinantes para un bar
+      col = "blue")
```

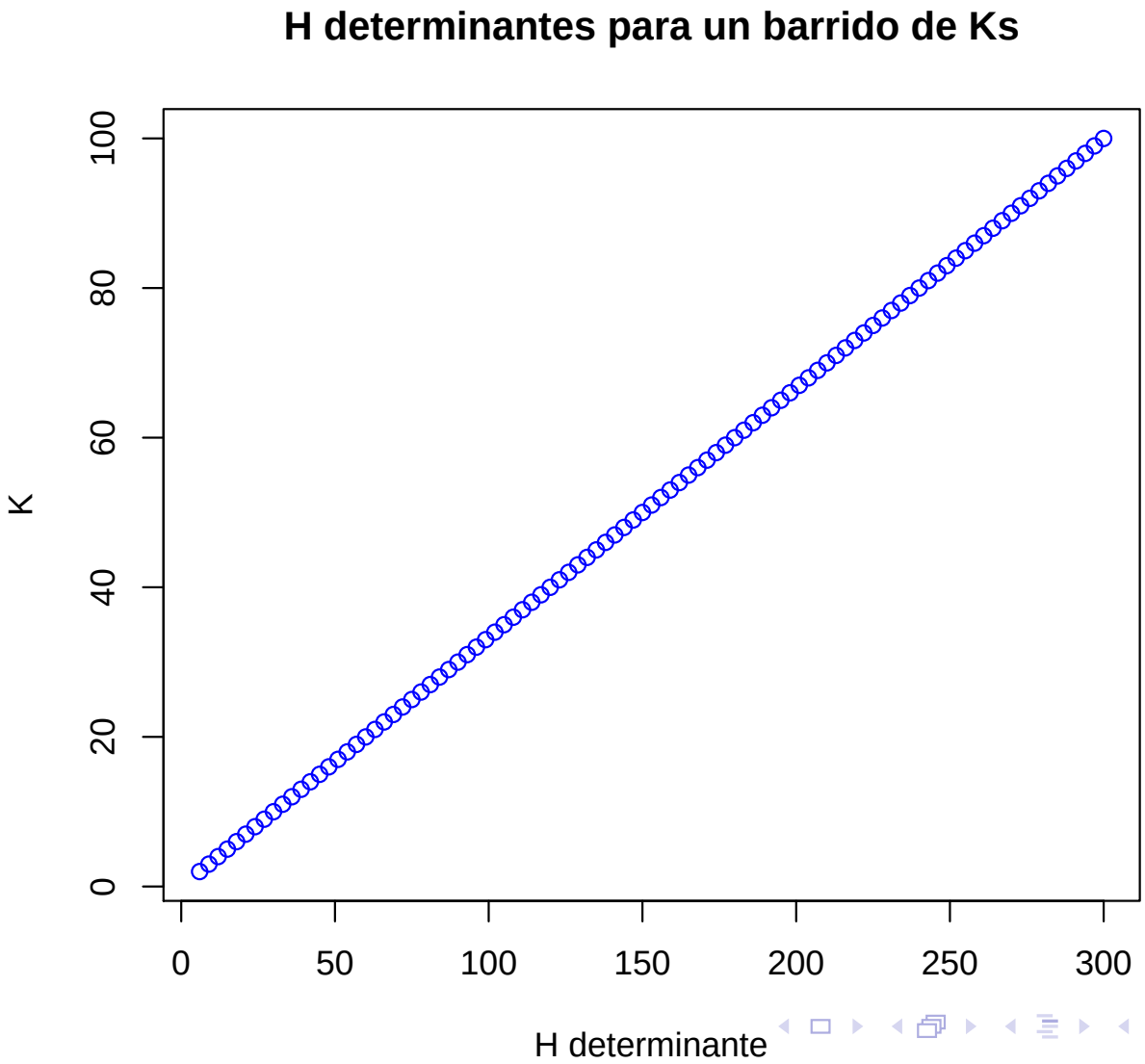
# Hs para K





# Hs determ. para K





# Conclusión



- Podemos concluir que con $P(k_i)$ iguales para todas las $k_i$ y $h = 3 * k$ tenemos 0.95 de probabilidad de que nuestros $k$ elementos aparezcan al menos 1 vez.
- ¿Por qué? :)