MODELOS PROBABILISTICOS Y ANALISIS DE DATOS

2020-2

Módulo 3: Inferencia estadística y métodos bayesianos

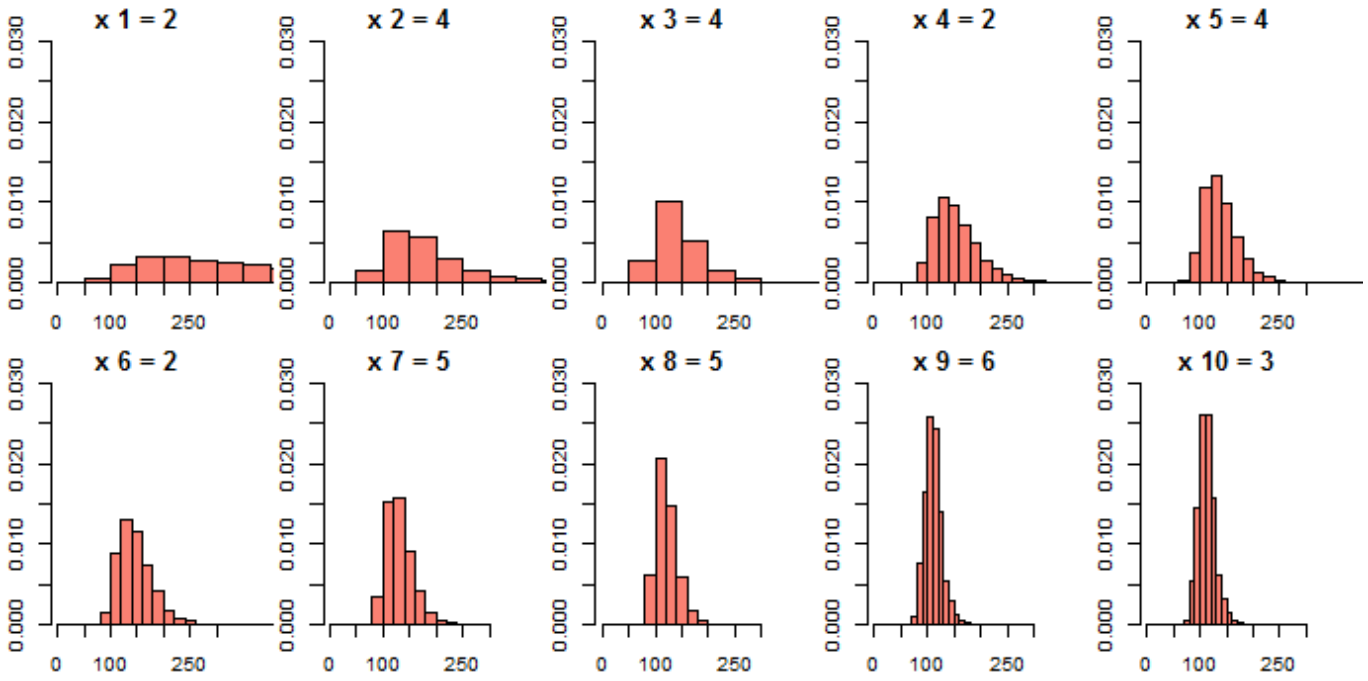
Javier Riascos Ochoa, PhD

Ejercicio 4: Inferencia Bayesiana, distribuciones a posteriori, intervalos de credibilidad y actualización bayesiana

En el ejercicio de *Capture/re-capture* suponga que el número real de peces en el lago es N_{real}, número que se quiere estimar utilizando la aproximación bayesiana aplicada a una serie de observaciones de los peces marcados recapturados en diferentes campañas. Como en la presentación, se realiza una campaña de captura y marcación de $m < N_{real}$ peces los cuales son devueltos al lago. Posteriormente, se realizan n_{sample} campañas de recaptura y conteo de peces marcados. El tiempo entre campañas es tal que los peces se alcanzan a mezclar, recapturándose k peces y observando el número de peces marcados X_i en cada campaña. Los peces recapturados no se vuelven tímidos. Observe que los X_i se distribuyen Hipergeométrico con parámetros (siguiendo la notación en **R**):

$$X_i \sim Hypergeometric(m, n = N_{real} - m, k)$$

1. **Simulación para el muestreo:** Suponga: $N_{real} = 100$, $m = 20$, $k = 20$. Utlizando la función `rhyper` en **R** genere $n_{sample} = 10$ realizaciones de X que simulan el número de peces marcados recapturados en 10 campañas. Llame a este vector `x`.
2. **Primera distribución a-posteriori:** Suponga la distribución a-priori uniforme entre 20 y 500 para el número de peces en el lago `n_fish`. Obtenga una primera distribución a-posteriori (`post_fish`) utilizando como dato de entrada el primer elemento del vector `x` (es decir X_1). Dibuje su histograma y calcule su media, desviación estándar e intervalo de credibilidad Bayesiana del 95% (a partir de los cuantiles 0.025, 0.975). Para este punto modifique la parte del código *"Capture Re-capture Bayesian.r"* que se muestra en las siguientes páginas.
3. **Actualización bayesiana:** Para las observaciones siguientes ($i = 2, \dots, 10$), utilice como distribución a-priori la distribución a-posteriori de la campaña anterior ($i - 1$). Modifique el código suministrado más adelante (que muestra sólo un segundo update) para realizar estos cálculos recursivamente y que muestre los histogramas de las distribuciones a-posteriori, medias, desviaciones estándar e intervalo de confianza del 95% para cada una de ellas. La salida debe ser similar a lo mostrado en la siguiente página.
4. **Analice**: ¿cómo varían las distribuciones a-posteriori con cada "update" de los datos?, ¿cómo varían sus medias, desviaciones estándar e intervalos de confianza? ¿se acerca la media al valor real N_{real} del número de peces en el lago?, ¿se encuentra N_{real} en los intervalos de confianza de las distribuciones?



	x	mean	sd	q0.025	q0.975
1	2	279.0013	110.44871	101	484
2	4	183.4085	77.30620	87	390
3	4	145.0421	47.51400	84	262
4	2	156.4048	46.02370	94	267
5	4	140.1406	33.82755	90	221
6	2	147.9000	33.32663	98	227
7	5	131.5025	25.13883	93	190
8	5	121.5150	20.26170	89	169
9	6	111.4624	16.20930	85	149
10	3	113.1499	15.68467	87	150

Código ejemplo para el punto 2

```
# The hypergeometric distribution is used
# as it implements the same process as the fish picking model.
# This code assumes that the number of recaptured marked fishes (n_marked) is 5

n_draw <- 100000

# Defining and drawing from the prior distribution (uniform)
n_fish <- sample(20:500, n_draw, replace = TRUE)

# Histogram
hist(n_fish)

# Defining the generative model and its simulation
# from the hypergeometric distribution
n_marked <- rep(NA, n_draw)
for(i in 1:n_draw) {
  n_marked[i] <- rhyper(1, m = 20, n=n_fish[i] -20, k=20)
}
n_marked

# Filtering out those parameter values that didn't result in the
# data that we actually observed
post_fish <- n_fish[n_marked == 5]
hist(post_fish)
length(post_fish)

# The posterior distribution showing the probability of different number of fish
# (binning here in bins of 20 just make the graph easier to interpret)
barplot(table(cut(post_fish, seq(0, 250, 20))) / length(post_fish), col = "salmon")
mean(post_fish)
sd(post_fish)
quantile(post_fish,c(0.025,0.975))
```

Código ejemplo para el punto 3

```
# Bayesian updating (actualización bayesiana)
# Two capture/re-capture experiments, 20 fishes initially captured
# and marked in each experiment, and no shy fishes:
# In all experiments 5 of 20 fishes re-captured were marked

# ----- 1st experiment (this is the same as the first code) -----

n_draw <- 100000

# Defining and drawing from the prior distribution (uniform)
n_fish <- sample(20:500, n_draw, replace = TRUE)

# Histogram
hist(n_fish)

# Defining the generative model and its simulation
# from the hypergeometric distribution
n_marked <- rep(NA, n_draw)
for(i in 1:n_draw) {
  n_marked[i] <- rhyper(1, m = 20, n=n_fish[i] -20, k=20)
}
n_marked

# Filtering out those parameter values that didn't result in the
# data that we actually observed
post_fish <- n_fish[n_marked == 5]
hist(post_fish)
length(post_fish)

# The posterior distribution showing the probability of different number of fish
# (binning here in bins of 20 just make the graph easier to interpret)
barplot(table(cut(post_fish, seq(0, 250, 20))) / length(post_fish), col = "salmon")
mean(post_fish)
sd(post_fish)
quantile(post_fish,c(0.025,0.975))

# ----- 2nd experiment (new code, but similar!) ----

# previous posterior is the new prior; sample from it
n_fish <- sample(post_fish, n_draw, replace = TRUE)
n_draw = length(n_fish)
hist(n_fish)

# Defining the generative model and its simulation
# from the hypergeometric distribution
n_marked <- rep(NA, n_draw)
for(i in 1:n_draw) {
  n_marked[i] <- rhyper(1, m = 20, n=n_fish[i] -20, k=20)
}
n_marked

# Filtering out those parameter values that didn't result in the
# data that we actually observed
post_fish <- n_fish[n_marked == x[2]]
length(post_fish)
hist(post_fish)
barplot(table(cut(post_fish, seq(0, 250, 20))) / length(post_fish), col = "salmon")
mean(post_fish)
sd(post_fish)
quantile(post_fish,c(0.025,0.975))
```