Final-2023

Mariano Villafuerte -156057 Mario Medina - 156940

30 noviembre 2023

Contents

1 Pruebas de hipótesis

1.1 Preparatoria en Illinois

1.1.1 Contexto

De acuerdo a una encuesta en EUA, 26% de los residentes adultos de Illinois han terminado la preparatoria. Un investigador sospecha que este porcentaje es menor en un condado particular del estado. Obtiene una muestra aleatoria de dicho condado y encuentra que 69 de 310 (22.26%) personas en la muestra han completado la preparatoria. Estos resultados soportan su hipótesis? (describe tu elección de prueba de hipótesis, valor p y conclusión).

Podemos tomar 2 enfoques, a continuación explicamos el porqué

- **Prueba con estadístico Z:** dado que hablamos de proporciones sabemos cuál es el error estándar de una proporción, podremos calcular el estadístico Z.

- **Enfoque bayesiano:** el 26% nos ayuda a definir una a priori y con los datos podemos generar una posterior. No es un cálculo de prueba de hipótesis tal cual pero podemos obtener intervalos de confianza que nos ayuden a determinar si realmente es significativamente menor.

Empezamos con la **prueba del estadístico Z**, nuestra prueba de hipótesis la podemos definir como (1 cola)

$$H_0 : \hat{\theta} = 0.26$$
$$H_1 : \hat{\theta} < 0.26$$

Y el estadístico se vería de la siguiente forma. Sabemos que $\hat{\theta} = \frac{69}{310}$

$$Z = \frac{\hat{\theta} - 0.26}{\sqrt{\frac{0.26(1-0.26)}{310}}} = -1.502016$$

El valor-p considerando que es de una cola sería, en específico la izquierda

$$p - value = P(Z < z)$$

el cálculo se ve de la siguiente manera.

```
numerador = (69/310)-0.26
denominador = sqrt((0.26*0.74)/310)
p_value <- pnorm(numerador/denominador)
print(paste0("El valor p asociado a esta prueba es: ", round(p_value,2)))
```

```
## [1] "El valor p asociado a esta prueba es: 0.07"
```

La conclusión es que no es significativo al 95% de confianza. Debido a que es mayor al valor crítico de 5%, por lo que no hay suficiente evidencia para rechazar la hipótesis nula.

Enfoque bayesiano: El problema trata de la estimación de una proporción, llamémosle θ donde θ es la proporción de adultos que terminaron la preparatoria en el condado específico de Illinois. Podemos asumir una a priori $P(\theta)$ que siga la información inicial que nos dice que ese porcentaje dentro de Illinois es de aproximadamente 26%, entonces usaremos una **Beta** que después de prueba y error tiene los parametros *(4,11)* que tiene de media 0.26 sin estar muy concentrada.

```
set.seed(156057)
sim_inicial <- tibble(theta = stats::rbeta(10000,4,11))
ggplot(sim_inicial) +
  geom_histogram(aes(x = theta, y = ..density..), bins = 15, color = "lightblue") +
  geom_vline(xintercept = 0.26, color = "red") +
  labs(title = "Distribución Inicial", subtitle = "Beta(4,11)") +
  xlab("Theta") + ylab("Densidad") +
  ggpubr::theme_pubclean(base_size = 12)
```

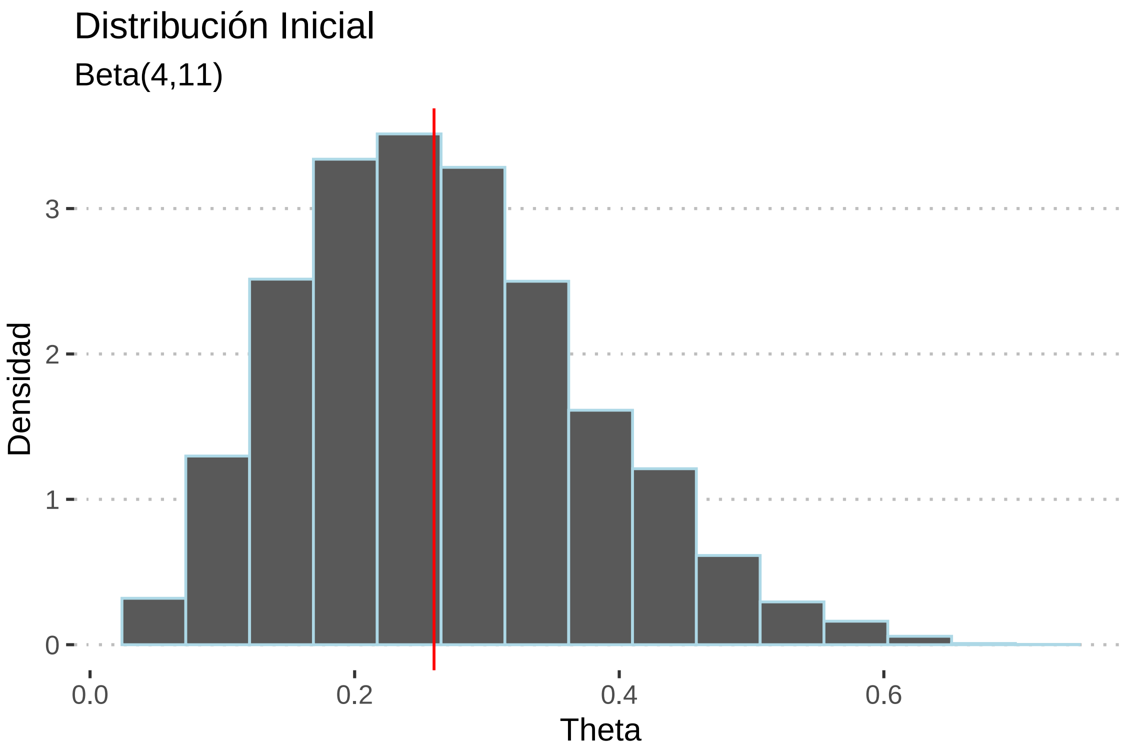

Figure 1: Distribución a priori: Beta (4,11)

Los datos del condado nos dicen que de 310 individuos únicamente tenemos 69 con preparatoria concluida (éxitos) entonces nuestra posterior queda de la siguiente manera

$$P(\theta|X) \propto P(X|\theta)P(\theta)$$
$$P(\theta|X) \propto \theta^{69+3}(1-\theta)^{241+10}$$
$$P(\theta|X) \propto \theta^{72}(1-\theta)^{251}$$

Obtenemos los siguientes histogramas.

```r
set.seed(156057)
sim_inicial <- sim_inicial %>% mutate(dist = "inicial")
sim_posterior <- tibble(theta = rbeta(10000, 73, 252)) %>%
  mutate(dist = "posterior")

sims <- bind_rows(sim_inicial, sim_posterior)

ggplot(sims, aes(x = theta, fill = dist)) +
  geom_histogram(aes(x = theta), bins = 30,
                 alpha = 0.5, position = "identity") +
  geom_vline(xintercept = 0.26, color = "red") +
  geom_vline(xintercept = 0.2246, color = "red") +
  labs(title = "Distribución Inicial y Posterior",
       subtitle = "Beta(4,11) vs. Beta(73,252)") +
  xlab("theta") + ylab("Densidad") +
  theme_pubclean(base_size = 12)
```

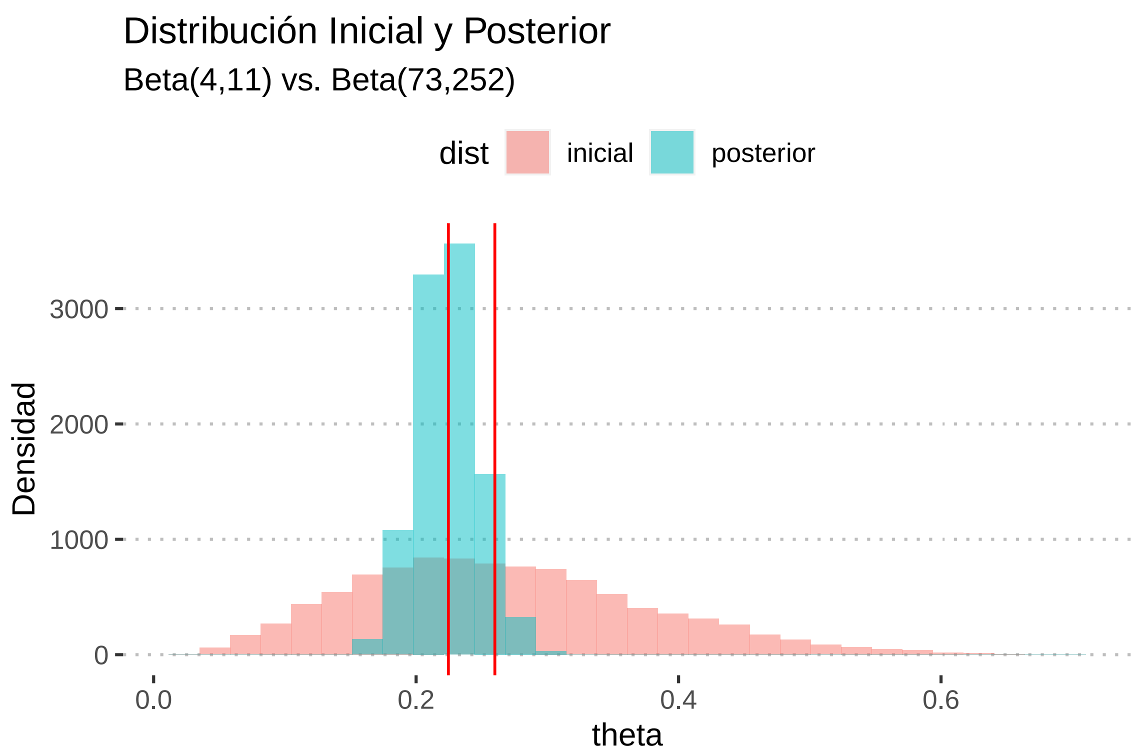

Figure 2: Inicial vs. Posterior

Un estimador puntual es la media de la distribuciones, que al ser una Beta se obtiene como $\frac{a}{a+b}$

- Dist. Inicial: $\frac{4}{4+11} = 0.267$
- Dist. Posterior: $\frac{73}{73+252} = 0.2246$
- Máxima verosimilitud: $\frac{69}{310} = 0.2226$

Asímismo podemos obtener intervalos de confianza fácilmente debido a nuestra distribución "conocida"

```
set.seed(156057)
paste0("Intervalo dist. Inicial: (",
       round(qbeta(0.025, shape1 = 4, shape2 = 11),2),",",
       round(qbeta(0.975, shape1 = 4, shape2 = 11),2),")")
```

```
## [1] "Intervalo dist. Inicial: (0.08,0.51)"
```

```
paste0("Intervalo dist. posterior: (",
       round(qbeta(0.025, shape1 = 73, shape2 = 252),2),",",
       round(qbeta(0.975, shape1 = 73, shape2 = 252),2),")")
```

```
## [1] "Intervalo dist. posterior: (0.18,0.27)"
```

Dado la información anterior el **intervalo de confianza al 95% de la posterior sí incluye el valor del 26%. Por lo que todavía no es del todo "aceptable"** que la proporción del condado sea significativamente menor al del estado de Illinois

La conclusión de este enfoque es el mismo que en el primero.

1.2 Chicaros de Mendel

1.2 Mendel criaba chícharos de semillas lisas amarillas y de semillas corrugadas verdes. Éstas daban lugar a 4 tipos de descendientes: amarrillas lisas, amarillas corrugadas, verdes lisas y verdes corrugadas. El número de cada una es multinomial con parámetro $p = (p_1, p_2, p_3, p_4)$. De acuerdo a su teoría de herencia este vector de probabilidades es:

$$p = (9/16, 3/16, 3/16, 1/16)$$

A lo largo de $n = 556$ experimentos observó $x = (315, 101, 108, 32)$. Utiliza la prueba de cociente de verosimilitudes para probar $H_0 : p = p_0$ contra $H_0 : p \neq p_0$.

1.3. Sean $X_1, ...X_n \sim Poisson(\lambda)$,

- Sea $\lambda_0 > 0$. ¿Cuál es la prueba Wald para $H_0 : \lambda = \lambda_0, H_1 : \lambda \neq \lambda_0$
- Si $\lambda_0 = 1$, $n = 20$ y $\alpha = 0.05$. Simula $X_1, ...X_n \sim Poisson(\lambda_0)$ y realiza la prueba Wald, repite 1000 veces y registra el porcentaje de veces que rechazas H_0, qué tan cerca te queda el error del tipo 1 de 0.05?