Entendiendo los valores-p

Paula Vargas Pellicer

10/03/2022

Calcular un valor p de una distribución normal

Vamos a usar datos inventados de la próxima consulta ciudadana sobre la revocación del mandato del actual presidente de los Estados Unidos Mexicanos

Te han asignado la tarea de realizar encuestas de salida en dos estados para determinar si la revocación del mandato procede o si el presidente sigue en la presidencia. Vamos a tomar Morelos y Guanajuato como ejemplos (por sus tendencias electorales pasadas).

```
library(tidyverse)

# Vamos primero a ver los resultados reales (inventados) del 10 de abril para
entender los resultados de la consulta


Morelos <- c(rep("sigue", 2473707), rep("revoca", 2461779))
Guanajuato <- c(rep("sigue", 153778), rep("revoca", 189951))

proportions(table(Morelos))

proportions(table(Guanajuato))
```

Un muestreo de 1,000 votantes

Empecemos por simular las encuestas de salida en Guanajuato

```
set.seed(2020)

survey_a1 <- sample(Guanajuato,1000,replace=F)

table(survey_a1) %>% as.data.frame %>%
  ggplot(aes(x=survey_a1, y = Freq, label = Freq)) +
  geom_col(fill=c("blue", "red")) +
  geom_label(label = paste(proportions(table(survey_a1))*100,"%")) +
  theme_minimal() +
  ggtitle("Resultados de encuesta en Guanajuato (N = 1,000)")
```

```r
prop.test(table(survey_a1))
```

¿Por qué no tenemos el valor exacto real en la encuesta?

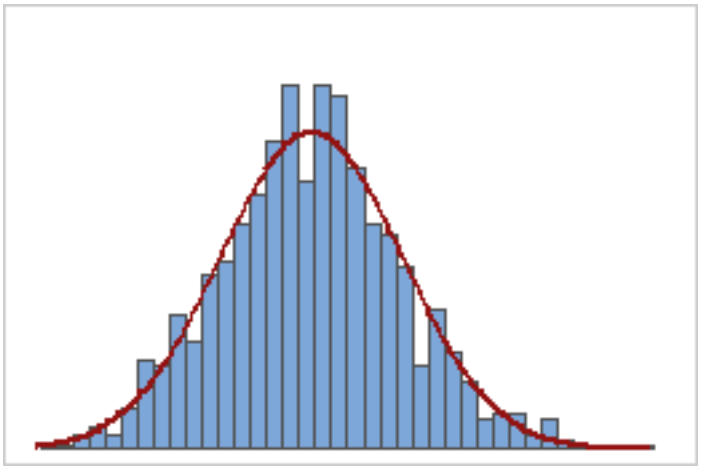
Sabemos que el verdadero porcentaje de votos a favor de que se quede en el mandato en Guanajuato fue del 44.7%. En nuestra encuesta de 1,000 votantes, fue del 43.6%, un poco más bajo. Primero establezcamos por qué el porcentaje en nuestra muestra no es exactamente igual al valor real, asumiendo que nuestros encuestados son completamente honestos (son simulados, después de todo).

Dado que extrajimos una muestra aleatoria de una población, nuestros resultados pueden diferir un poco de los resultados verdaderos como resultado del azar. Es posible que seleccionemos algunos votantes en contra de más, o algunos votantes a favor de más en nuestra muestra aleatoria, en comparación con el porcentaje real.

Esto es lo que plantea el llamado Teorema del Límite Central (CLT) en estadística:

Si repites un experimento una gran cantidad de veces, tu estadística de resumen (p. ej., valor medio de conteo de votos en contra de revocación) se dispersará alrededor del valor real esperado "verdadero". A veces estará por debajo del valor esperado, a veces por encima. Pero con mayor frecuencia estará más cerca del valor real, y con menos frecuencia estará muy lejos.

De hecho, si repites el experimento muchas veces, tus resultados se distribuirán en forma de una curva de campana (distribución normal) alrededor del verdadero valor esperado. Otra cosa importante para recordar: cuanto más a menudo repitas el experimento, más cerca estarás del valor real si promedias todos los resultados de todos los experimentos. Esto describe la distribución normal de los datos que ya habíamos visto:



Es por ello, que con muestras suficientemente grandes, las pruebas de normalidad de datos son menos “necesarias”.

Seleccionamos al azar a 1000 votantes y calculamos la proporción de votos en contra de la revocación. Nuestro resultado podría estar un poco por debajo o un poco por encima del verdadero resultado real de la elección como resultado de una probabilidad aleatoria

durante el proceso de muestreo. Sin embargo, es más probable que estemos cerca del valor real que lejos. Si pudiéramos repetir esto muy a menudo, por ejemplo, seleccionamos otros 1,000 votantes y luego otro, y hacemos esto 1,000 veces y anotamos la proporción de votos en contra de la revocación en cada una de nuestras encuestas, obtendríamos una distribución normal alrededor del valor real. Si pudiéramos repetir la encuesta 100,000 veces en lugar de solo 1,000, sería más probable que termináramos mucho más cerca del valor real y menos probable que sobrestimáramos o subestimáramos aleatoriamente el valor real por un amplio margen.

Repetición de la encuesta

Si tuvieras tiempo y recursos infinitos, podrías realizar una segunda encuesta, ya que tu primera encuesta indicó que la revocación ganó Guanajuato, pero quieres asegurarte de que no estás, por casualidad, un poco por encima del verdadero resultado y, en realidad, "que se quede AMLO" ganó (nuevamente, ten en cuenta que no sabrías el verdadero resultado en el momento de la encuesta de salida, que es la razón principal para realizar la encuesta de salida).

Una segunda encuesta:

```
survey_a2 <- sample(Guanajuato,1000,replace=F)

table(survey_a2) %>% as.data.frame %>%
  ggplot(aes(x=survey_a2, y = Freq, label = Freq)) +
  geom_col(fill=c("blue", "red")) +
  geom_label(label = paste(proportions(table(survey_a2))*100,"%")) +
  theme_minimal() +
  ggtitle("Resultados de la segunda encuesta en Guanajuato (N = 1,000)")
```

En tu segunda encuesta, también gana la revocación. Esta vez, el porcentaje de votos es un poco más bajo, pero aún lidera por un amplio margen.

Pero realmente quieres saber la verdad y, por lo tanto, recurres a una medida extrema: realizas 1,000 encuestas de 1,000 votantes cada una y anotas el porcentaje de votos de revocación vs. seguir en el mandato en cada encuesta, y al final cuentas cuántas de las encuestas muestran a la revocación a la cabeza.

```
draw_sample <- function(x,n=1000){
  s <- sample(x, n, F)
  proportions(table(s))
}

set.seed(2020)

Guanajuato_sim <- replicate(1000,draw_sample(Guanajuato))
```

Este código crea una función que brinda una muestra de 1,000 votantes y devuelve los porcentajes de votos en contra y a favor. Luego ejecutamos (replicamos) esta función 1,000

veces y almacenamos los resultados en el objeto `Guanajuato_sim`. Mira los primeros dos resultados de nuestro meta-estudio:

```
Guanajuato_sim[1:2,1:20]
```

Los primeros 20 resultados de 1,000 encuestas simuladas con 1,000 votantes cada muestran una imagen muy clara. Vemos que la revocación siempre lidera por un amplio margen. A veces es 54%-46%, a veces es 58%-42%. Entonces, hay cierta variabilidad, pero nunca vemos "que siga" a la cabeza en estas encuestas.

Tracemos los resultados de todas nuestras 1,000 encuestas:

```
dat <- data.frame(t(Guanajuato_sim))

qplot() + theme_minimal() +
  geom_histogram(aes(x=dat$sigue),fill="red", alpha=.5) +
  geom_histogram(aes(x=dat$revoca), fill="blue", alpha=.5) +
  geom_vline(xintercept = mean(dat$sigue), color="red") +
  geom_vline(xintercept = mean(dat$revoca), color="blue") +
  ggtitle("1,000 encuestas en Guanajuato de 1,000 votantes cada una ",
          subtitle="Histogramas de valores simulados para revocación (azul) y
seguir en mandato (rojo)")

c(mean(dat$sigue), mean(dat$revoca))
```

Si promediamos todas nuestras 1,000 encuestas (líneas verticales rojas y azules en el centro), estamos más cerca del valor verdadero real.

```
c(mean(dat$sigue), mean(dat$revoca))
```

Fíjate y compara con los resultados reales (¡inventados!) de la encuesta.

Si pudiéramos repetir un experimento (como una encuesta entre una muestra aleatoria de votantes) muchas veces, el valor promedio de todos estos experimentos estaría muy cerca del valor real. Sin embargo, un solo resultado de una sola encuesta puede ser algo diferente del verdadero resultado. Hay una pequeña posibilidad de que incluso podría estar muy lejos.

¿Cómo sabemos si una encuesta está cercana a la realidad?

Sabemos que el resultado de nuestra encuesta (proporción de votos a favor de la revocación) se encuentra en algún lugar de una distribución normal alrededor del valor real (como postula el teorema del límite central y como se muestra en nuestra simulación anterior). A partir de la forma y las propiedades bien conocidas de una distribución normal, podemos derivar la probabilidad de que nuestro resultado del 56% de los votos a favor de la revocación (en nuestra primera encuesta) esté, digamos, a 7 puntos porcentuales del valor real (lo que significa que en realidad los seguidores de AMLO podrían estar a la cabeza).

Hay dos parámetros que determinan el "ancho" de la distribución normal:

El número de observaciones (en nuestro caso, 1,000 votantes), y la desviación estándar (es decir, cuán grande es la variación en nuestros datos). Para una distribución binomial (donde solo hay dos resultados, a favor y en contra), la desviación estándar viene dada por la raíz cuadrada del porcentaje de una de las respuestas (favor de la revocación) multiplicada por la otra respuesta (que se quede) y el número de encuestados: 'sqrt(p*q*n)', donde p = porcentaje de revocación, q = porcentaje de sigue, n = número de observaciones. Aquí puedes ver una distribución normal con N = 1,000 alrededor del voto de Guanajuato (de nuestra primera encuesta):

```
voto <- .564

N <- 1000

mean <- voto*N

sd <- sqrt(voto*(1-voto)*N)

x <- (mean - 5*sd):(mean + 5*sd)

norm1 <- dnorm(x,mean,sd)

p <- pnorm(x,mean,sd)

qplot() + theme_minimal() +
  geom_line(aes(x=x,y=norm1), color="blue") +
  geom_area(aes(x=x,y=norm1),fill="blue",alpha=.1) +
  geom_vline(xintercept = x[which.min(abs(p-0.025))], color="blue") +
  geom_vline(xintercept = x[which.min(abs(p-0.975))], color="blue") +
  ggtitle("Distribucion normal con N = 1,000",
          subtitle= paste0("Lineas verticales = 95% intervalo alrededor de la
media: ",
                           round(x[which.min(abs(p-0.025))]/N*100,1), "% a ",
                           round(x[which.min(abs(p-0.975))]/N*100,1),"%"))
```

En la gráfica has marcado con las dos líneas verticales los límites del llamado intervalo del 95%. Esta es la zona donde se localizan el 95% de los casos. Afuera (cola izquierda o cola derecha) se juntan solo el 5% de las observaciones. Esto es importante porque, como es costumbre, usamos el intervalo del 95% para indicar los valores que aún consideramos "probables".

Si la proporción de votos a favor de revocación en Guanajuato fuera del 56,4% (como en nuestra primera encuesta), repitiendo una encuesta muchas veces con N = 1,000 votantes, el 95% de nuestros resultados oscilarían entre el 53.4% y el 59.5%.

Esto se llama el intervalo de confianza del 95%. Podríamos escribir: "Nuestra encuesta de salida sitúa la revocación del mandato con el 56.4% de los votos (intervalo de confianza del 95%: 53.4%-59.5%)".

Otra forma de expresarlo es decir que el margen de error es “+- 3.1%” (porque ésta es la distancia desde el valor más alto y más bajo del intervalo de confianza hasta nuestro resultado obtenido, por ejemplo, 59.5%-56.4%).

El mensaje importante de este análisis es: Nuestro intervalo de confianza está muy lejos de la marca del 50%. Es decir, podemos estar muy seguros de que la revocación del mandato está realmente a la cabeza sobre la decisión de que siga en Guanajuato. Sí, todavía hay cierta incertidumbre; el intervalo del 95% tiene más de seis puntos porcentuales de ancho, por lo que no podemos decir con seguridad si revocación en realidad obtuvo el 54% o más bien el 59% de los votos, pero podemos estar muy seguros de que "siga" no ha recibido tantos o más votos.

El efecto de tamaño de muestra

Como postula el teorema del límite central y la ley de los números grandes (y de la intuición), un tamaño de muestra más grande conducirá a un margen de error más pequeño. Veamos qué tan grande es el intervalo de confianza con una muestra de 10,000 votantes en lugar de 1,000 votantes:

```
N <- 10000
mean <- voto*N
sd <- sqrt(voto*(1-voto)*N)
x <- x*10
norm1 <- dnorm(x,mean,sd)
p <- pnorm(x,mean,sd)
qplot() + theme_minimal() +
  geom_line(aes(x=x,y=norm1), color="blue") +
  geom_area(aes(x=x,y=norm1),fill="blue",alpha=.1) +
  geom_vline(xintercept = x[which.min(abs(p-0.025))], color="blue") +
  geom_vline(xintercept = x[which.min(abs(p-0.975))], color="blue") +
  ggtitle("Distribucion normal con N = 1,000",
          subtitle= paste0("Lineas verticales = 95% intervalo alrededor de la
media: ",
                           round(x[which.min(abs(p-0.025))]/N*100,1), "% a ",
                           round(x[which.min(abs(p-0.975))]/N*100,1),"%"))
```

Como puedes ver, la distribución es mucho más estrecha. El margen de error es mucho menor, en lugar de un 56.4% más o menos un 3.1%, ahora es del 56.4%, más o menos un 1.1%. Lo que obviamente es una gran mejora. Pero en realidad también sería mucho más costoso encuestar 10,000 votantes. Por lo tanto, existe una compensación entre obtener resultados rápidos y económicos y obtener resultados que sean lo más precisos posible (Esto aplica, sobretodo, en la ecología).

Ten en cuenta cuando leas algo como "el 43% de las personas quiere X mientras que solo el 41% de las personas favorecen Y", según una encuesta de 1,000 personas, porque ahora sabes qué tan grande suele ser el margen de error para este tamaño de muestra.

¿Qué es el error estándar?

¿Puedes ver que al aumentar el número de observaciones conduce a una distribución más estrecha? a partir de la fórmula de la desviación estándar de una distribución binomial como se mencionó anteriormente, que es: sqrt(p*q*n) [donde p = porcentaje revocación, q = porcentaje de sigue, n = número de observaciones]. Entonces, si n aumenta por un factor de 100, entonces la desviación estándar sólo aumenta por el factor sqrt (100) = 10. Entonces, el tamaño relativo de la desviación estándar en proporción a su número de votantes por revocación ha disminuido.

La desviación estándar de la distribución de todos los resultados (hipotéticos) de nuestra encuesta también se denomina error estándar (SE) de nuestra estimación. Esto es importante porque mucha gente confunde la desviación estándar con el error estándar.

La desviación estándar cuantifica la variación dentro de un conjunto de medidas; nos dice cuánta variación hay en nuestros datos.

El error estándar cuantifica la variación de las medias de varios conjuntos de mediciones.

Podemos calcular fácilmente el error estándar solo a partir de la desviación estándar de la distribución de todos nuestros experimentos repetidos.

```
dat <- data.frame(t(Guanajuato_sim))
sd(dat$revoca)
```

Pero, si no tenemos miles de experimentos repetidos sino solo una encuesta real, tenemos que estimar el error estándar en función de nuestros datos existentes. La estimación del error estándar para una distribución binomial de una muestra viene dada por la fórmula: 'sqrt(p*(1-p)/n)', donde p es la proporción de votos de revocación y n es el tamaño de la muestra.

```
#Error estándar
sqrt(0.564*0.436/1000)
```

Error estándar: estimado a partir de una sola muestra en la que la proporción de votos de revocación fue del 56.4%. Los valores son muy similares. Es importante que sepamos esto porque significa que podemos inferir el error estándar de una sola encuesta, sin tener que simular o realizar miles.

Una regla general importante sobre la distribución normal, es que los límites del intervalo de confianza del 95% son aproximadamente: `valor medio +/- 2* error estándar`. Esto significa que si calculas el error estándar de la estimación de tu encuesta, puedes derivar el margen de error simplemente multiplicándolo por 2. (Para ser aún un poco más precisos, para la distribución t), la fórmula es intervalo de confianza del 9 % = media + /- 1,96*SE).

```
#Intervalo de confianza
se = sqrt(0.564*0.436/1000)
c(.546-2*se, .564+2*se)
```

Y a todo esto, ¿cómo lo hago en R?

```
binom.test(564,1000, p=0.5, alternative="greater")
```

Una encuesta en Morelos

```
set.seed(2020)

survey_g1 <- sample(Morelos,1000,replace=F)

table(survey_g1) %>% as.data.frame %>%
  ggplot(aes(x=survey_g1, y = Freq, label = Freq)) +
  geom_col(fill=c("red", "blue")) +
  geom_label(label = paste(proportions(table(survey_g1))*100,"%")) +
  theme_minimal() +
  ggtitle("Resultados de encuesta en Morelos (N = 1,000)")

prop.test(table(survey_g1))
```

Nuestra encuesta encuentra que la No-revocación está a la cabeza con el 50.7% de los votos, pero la diferencia entre revocación y no revocación no es estadísticamente significativa. La prueba nos da un valor p de 0.681, lo que significa: si "revoca" y "sigue" tuvieran la misma proporción de votos, la probabilidad de extraer una muestra aleatoria como la nuestra y encontrar que una de las opciones lidera con el 50.7% (o incluso más votos) sería del 68.1%. Por lo tanto, existe un alto riesgo de que un resultado como este pueda producirse por azar. Por lo tanto, no podemos rechazar la hipótesis nula.

¿Cómo evitamos un error de tipo II (dado por tamaño de muestra)?

¿Qué tan grande debe de ser mi muestra?

Aumentamos el tamaño de muestra y le preguntamos a 100,000 personas en vez de 1,000

```
set.seed(2020)

survey_g2 <- sample(Morelos,100000,replace=F)

table(survey_g2) %>% as.data.frame %>%
  ggplot(aes(x=survey_g2, y = Freq, label = Freq)) +
  geom_col(fill=c("red", "blue")) +
  geom_label(label = paste(proportions(table(survey_g2))*100,"%")) +
  theme_minimal() +
  ggtitle("Resultados de la segunda encuesta en Morelos (N = 100,000)")

prop.test(table(survey_g2))
```

Ahora estamos mucho más cerca del resultado electoral real con nuestra estimación de muestra de 50.07%. Sin embargo, de nuevo, nuestro resultado no es estadísticamente significativo. Entonces, incluso con esta muestra que es 100 veces más grande, nuestro resultado se encuentra dentro del margen de error y no podemos estar seguros del resultado observado.

Tenemos fondos ilimitados, por lo que decidimos preguntarle a 3 millones de personas (¡tan grandes son nuestros fondos que nos permite preguntarle a más personas que la población completa de Morelos!):

```
survey_g4 <- sample(Morelos,3000000,replace=F)

table(survey_g4) %>% as.data.frame %>%
  ggplot(aes(x=survey_g4, y = Freq, label = Freq)) +
  geom_col(fill=c("red", "blue")) +
  geom_label(label = paste(round(proportions(table(survey_g4))*100,4),"%")) +
  theme_minimal() +
  ggtitle("Resultados de la tercera encuesta en Morelos (N = 3,000,000)")

prop.test(table(survey_g4))
```

Finalmente, nuestro resultado es estadísticamente significativo y podemos estar bastante seguros de que "sigue" realmente ganó en Morelos. Pero tuvimos que pedirle a más de la población entera para obtener este resultado, por lo que nuestra "encuesta de salida" en este punto no es un esfuerzo menor en comparación con solo contar todos los votos reales.

Lo que quiero demostrar es que un efecto pequeño en una población (lo que sucede en Morelos) puede ser difícil de detectar con una muestra, lo que requiere un tamaño de muestra grande. Un gran efecto (como el caso de Guanajuato) puede detectarse con muestras más pequeñas.