Trabajo práctico N°4

Inferencia Estadística.

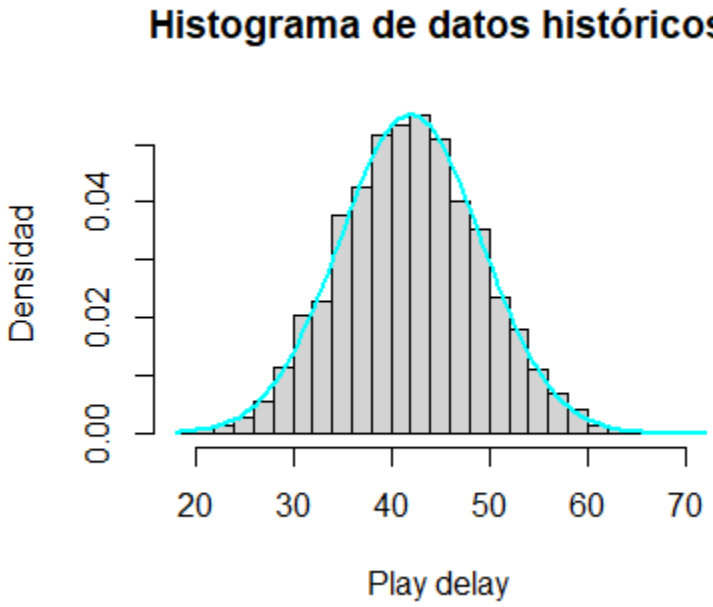
Timo Gabriel Doeswijk, Bautista Goya, José Nicolas Rodriguez López.

```
setwd("D:/Documents/- UTDT/inferencia/tp4") # Cambiar el argumento por la working
directory propia

# 1)
datos_historicos = read.csv("datos_historicos.csv")
play_delay_historico = datos_historicos$play.delay # Columna play.delay de los datos
historicos

# a)
mean(play_delay_historico) # Media muestral
var(play_delay_historico) # Varianza muestral

# b)
hist(play_delay_historico, freq = FALSE, breaks = 20) # Histograma de los datos
curve(dnorm(x, mean = mean(play_delay_historico), sd = sd(play_delay_historico)), col =
"cyan", lwd = 2, add = TRUE) # Normal con los datos muestrales

# 2)
datos_nuevos = read.csv("datos_nueva_version.csv")
play_delay_nuevo = datos_nuevos$play.delay # Columna play.delay de los datos nuevos

# a)
mean(play_delay_nuevo)

alpha <- seq(0.01,0.1,0.01)
zalpha <- sapply(alpha, qnorm)
zalpha <- -zalpha
Xprom <- sapply(zalpha, region)
```

1 - Validación visual de los datos históricos

a) (Leo la consigna) Mediante un código simple en R, obtuvimos, de los datos, que $\mu_{0,\,obs} = 41,98465$ y $\sigma^2_{0,\,obs} = 52,43255$

b) (Leo la consigna) Sí, los datos parecen distribuirse de forma normal ya que la forma del histograma se asemeja a la normal que se encuentra por encima (misma esperanza y desvío estándar).



2 - Grupo de prueba (nueva versión)

(Leo la consigna) Sabemos que el promedio es un buen estimador de la esperanza (μ). Por lo tanto:

$$\hat{\mu} = \overline{X}_{200} \Rightarrow \hat{\mu}_{obs} = \overline{X}_{200,obs} = 42.9895$$

3 - Construcción del test de hipótesis

a)

➔ $H_0: \mu_0 = \mu$, es decir, no existe diferencia en el “play-delay” con el cambio de software en la muestra observada.

➔ $H_1: \mu_0 < \mu$, es decir, la media del “play-delay” de la nueva versión es mayor a la de la versión anterior.

Justificación: Se busca rechazar actualizaciones que aumenten el “play-delay”, esto es equivalente a decir que se quiere demostrar que una nueva actualización aumenta la media de éste.

$P(EI) = P(Rechazar\ H_0 | H_0\ V)$ =“decidir que la actualización aumenta el ‘play-delay’ cuando en realidad esto no es así”

b)

En este caso, $P(EI) = P(Rechazar\ H_0 | H_0\ V) = \alpha = 0,05$.

Tenemos

$X_i = "'play - delay' del\ usuario\ i" \sim N(\mu_1, \sigma_0^2),\ 1 \leq i \leq 200,\ m.a\ i.i.d.$,

con μ desconocida y σ_0^2 conocida, queremos decidir entre $H_0: \mu_0 = \mu$ y $H_1: \mu_0 < \mu$.

Sabemos que $\overline{X}_{200} \sim N(\mu;\ \frac{\sigma_0^2}{n})$. Estandarizando, $T = \frac{\overline{X}-\mu}{\sqrt{\frac{\sigma_0^2}{n}}} \sim N(0,1)$ será nuestro estadístico.

Rechazar H_0 equivale a que el estadístico sea mayor a z_α(valor que deja área α a derecha). Entonces, $\alpha = P(T > z_\alpha | \mu_0 = \mu)$, es decir:

$0,05 = P(\frac{\overline{X}-\mu}{\sqrt{\frac{\sigma_0^2}{n}}} > z_{0,05} | \mu_0 = \mu) = P(\frac{\overline{X}-\mu_0}{\sqrt{\frac{\sigma_0^2}{n}}} > z_{0,05})$ (Bajo H_0).

Por tabla, $z_{0,05} = 1,645$, por lo tanto buscamos $P(\frac{\overline{X}-\mu_0}{\sqrt{\frac{\sigma_0^2}{n}}} > 1,645)$ (Bajo H_0).

Por datos, tenemos que $\mu_0 = 41,98465,\ n = 200$ y $\sigma_0^2 = 52,43255$

Por lo tanto, nuestra región de rechazo es $R = \left\{ \frac{\overline{X}-\mu_0}{\sqrt{\frac{\sigma_0^2}{n}}} > 1,645 \right\} \approx$

$\left\{ \frac{\overline{X}-41,98465}{0,5120183} > 1,645 \right\} = \{\overline{X} > 42,82692\}$

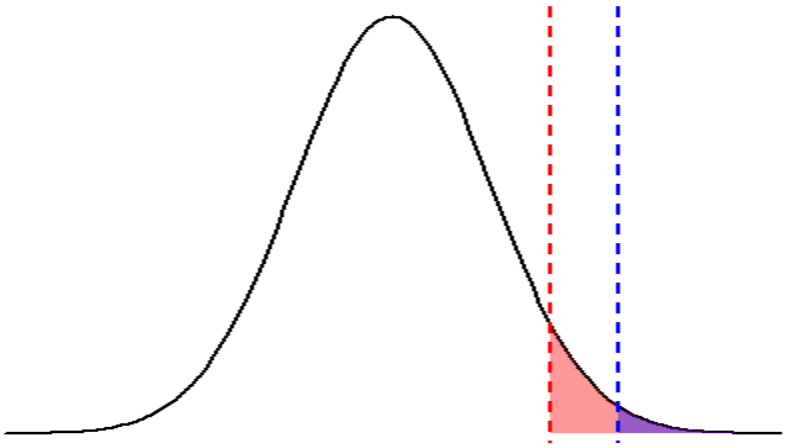
4- Toma de decisiones

a)

Basándose en las nuevas 200 observaciones, obtenemos que $\overline{X}_{obs} = 42,9895 > 42,82692$, por lo que **SI** se envía el código a revisión. Es decir, rechazo H_0.

b)

Con un nivel $\alpha_1 = 0,01$, tenemos $z_{0,01} = 2,33$. Entonces, la región de rechazo es $R' = \left\{\frac{\overline{X}-41,98465}{0.5120183} > 2,33\right\} = \{\overline{X} > 43,17765\}$. R' requiere un mayor promedio muestral para rechazar H_0 que R, es decir, $R' < R$.



En colorado, $z_{0,05}$ y, en azul, $z_{0,01}$

En general, niveles de significancia menores implicarán regiones de rechazo menores, ya que el nivel indica con qué probabilidad ocurre un error de tipo uno y, al ser un nivel menor, se necesitará más evidencia (un estadístico observado más grande) para rechazar H_0 y, en consecuencia, cometer un error de tipo uno. Se puede ver en las áreas de las gaussianas para los distintos niveles.

Para niveles de significancia mayores, la probabilidad de que ocurra un error de tipo uno será mayor ya que se necesitará menos evidencia para rechazar H_0, y por lo tanto la región de rechazo será mayor.

c) Considerando el conjunto $\alpha = \{0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1\}$,

➔ $\alpha = 0,01$

$\Rightarrow z_{0,01} = 2,326348 \Rightarrow R' = \left\{\frac{\overline{X}-41,98465}{0.5120183} > 2,326348\right\} = \{\overline{X} > 43,17765\} \Rightarrow \overline{X}_{obs} = 42,9895 < 43,17765$
$\Rightarrow No\ rechazo\ H_0$

➔ $\alpha = 0,02$

$\Rightarrow z_{0,02} = 2,053749 \Rightarrow R' = \left\{\frac{\overline{X}-41,98465}{0.5120183} > 2,053749\right\} = \{\overline{X} > 43,03620\} \Rightarrow \overline{X}_{obs} = 42,9895 < 43,03620$
$\Rightarrow No\ rechazo\ H_0$

➔ $\alpha = 0,03$

$\Rightarrow z_{0,03} = 1,880794 \Rightarrow R' = \left\{\frac{\overline{X}-41,98465}{0.5120183} > 1,880794\right\} = \{\overline{X} > 42,94765\} \Rightarrow \overline{X}_{obs} = 42,9895 > 42,94765$
$\Rightarrow Rechazo\ H_0$

➔ $\alpha = 0,04$

$\Rightarrow z_{0,04} = 1,750686 \Rightarrow R' = \left\{\frac{\overline{X}-41,98465}{0.5120183} > 1,750686\right\} = \{\overline{X} > 42,88103\} \Rightarrow \overline{X}_{obs} = 42,9895 > 42,88103$
$\Rightarrow Rechazo\ H_0$

➔ $\alpha = 0,05$

$\Rightarrow z_{0,05} = 1,644854 \Rightarrow R' = \left\{\frac{\overline{X}-41,98465}{0.5120183} > 1,644854\right\} = \{\overline{X} > 42,82685\} \Rightarrow \overline{X}_{obs} = 42,9895 > 42,82685$
$\Rightarrow Rechazo\ H_0$

➔ $\alpha = 0,06$

$\Rightarrow z_{0,06} = 1,554774 \Rightarrow R' = \left\{\frac{\overline{X}-41,98465}{0.5120183} > 1,554774\right\} = \{\overline{X} > 42.78072\} \Rightarrow \overline{X}_{obs} = 42,9895 > 42.78072$
$\Rightarrow Rechazo\ H_0$

➔ $\alpha = 0,07$

$\Rightarrow z_{0,07} = 1,475791 \Rightarrow R' = \left\{\frac{\overline{X}-41,98465}{0.5120183} > 1,475791\right\} = \{\overline{X} > 42,74028\} \Rightarrow \overline{X}_{obs} = 42,9895 > 42,74028$
$\Rightarrow Rechazo\ H_0$

➔ $\alpha = 0,08$

$\Rightarrow z_{0,08} = 1,405072 \Rightarrow R' = \left\{\frac{\overline{X}-41,98465}{0.5120183} > 1,405072\right\} = \{\overline{X} > 42,70407\} \Rightarrow \overline{X}_{obs} = 42,9895 > 42,70407$
$\Rightarrow Rechazo\ H_0$

➔ $\alpha = 0,09$

$\Rightarrow z_{0,09} = 1,340755 \Rightarrow R' = \left\{\frac{\overline{X}-41,98465}{0.5120183} > 1,340755\right\} = \{\overline{X} > 42,67114\} \Rightarrow \overline{X}_{obs} = 42,9895 > 42,67114$
$\Rightarrow Rechazo\ H_0$

➔ $\alpha = 0,1$

$\Rightarrow z_{0,1} = 1,281552 \Rightarrow R' = \left\{\frac{\overline{X}-41,98465}{0.5120183} > 1,281552\right\} = \{\overline{X} > 42,64083\} \Rightarrow \overline{X}_{obs} = 42,9895 > 42,64083$
$\Rightarrow Rechazo\ H_0$

$\alpha_1 = 0.03$ será el elemento del conjunto más chico para el cual rechazo H_0.

α	z_α	R'	$R = \{\overline{X} > \ ...\}$	$Rechazo$

$0,01$	$2,326348$	$\left\{\frac{\overline{X}-41,98465}{0.5120183} > 2,326348\right\}$	$43,17578$	*No*
$0,02$	$2,053749$	$\left\{\frac{\overline{X}-41,98465}{0.5120183} > 2,053749\right\}$	$43,03621$	*No*
$0,03$	$1,880794$	$\left\{\frac{\overline{X}-41,98465}{0.5120183} > 1,880794\right\}$	$42,94765$	*Sí*
$0,04$	$1,750686$	$\left\{\frac{\overline{X}-41,98465}{0.5120183} > 1,750686\right\}$	$42,88103$	*Sí*
$0,05$	$1,644854$	$\left\{\frac{\overline{X}-41,98465}{0.5120183} > 1,644854\right\}$	$42,82685$	*Sí*
$0,06$	$1,554774$	$\left\{\frac{\overline{X}-41,98465}{0.5120183} > 1,554774\right\}$	$42,78072$	*Sí*
$0,07$	$1,475791$	$\left\{\frac{\overline{X}-41,98465}{0.5120183} > 1,475791\right\}$	$42,74028$	*Sí*
$0,08$	$1,405072$	$\left\{\frac{\overline{X}-41,98465}{0.5120183} > 1,405072\right\}$	$42,70407$	*Sí*
$0,09$	$1,340755$	$\left\{\frac{\overline{X}-41,98465}{0.5120183} > 1,340755\right\}$	$42,67114$	*Sí*
$0,1$	$1,281552$	$\left\{\frac{\overline{X}-41,98465}{0.5120183} > 1,281552\right\}$	$42,64083$	*Sí*

d) En el siguiente archivo, se encuentra la grilla de significancia tomando 3 decimales en vez de 2: tabla_resultados.pdf. $\alpha_1 = 0.025$ será el nivel de significancia más bajo para el cuál rechazo H_0.

5 - Significancia y error de tipo 1

a)
Por lo planteado en el ejercicio 3-b, sabemos que la región de rechazo con nivel $\alpha = 0,05$ bajo H_0 es:

$$R = \left\{ \frac{\overline{X}-\mu_0}{\sqrt{\frac{\sigma_0^2}{n}}} > 1,645 \right\}$$

Reemplazamos los datos y construimos la región de rechazo:

$$\mu_0 = 41,98465,\ n = 200 \text{ y } \sigma_0^2 = 52.43255$$

$$R = \left\{ \frac{\overline{X} - \mu_0}{\sqrt{\frac{\sigma_0^2}{n}}} > 1,645 \right\} \approx \left\{ \frac{\overline{X} - 41,98465}{0,5120183} > 1,645 \right\} = \{\overline{X} > 42,82692\}$$

Usando la simulación, obtenemos que:

$$\overline{X}_{obs} = 42,95907 > 42,82692$$

Por lo tanto, rechazamos la hipótesis nula. Sin embargo, esta NO es la decisión correcta ya que estamos cometiendo un error de tipo 1 al rechazar la hipótesis nula siendo ésta verdadera.

b)

En este experimento las muestras se simulan bajo la hipótesis nula, es decir, suponiendo que H_0 es verdadera. El nivel de significancia α = 0.05 representa la probabilidad de cometer un error tipo I (rechazar H_0 cuando en realidad es cierta). Por lo tanto, el test debería tomar la decisión correcta (no rechazar H_0) aproximadamente el 95% de las veces.

En la simulación, se toma la decisión correcta 9466 de 10000 veces, lo cual es acorde a lo esperado, ya que representa un 94,66% de las simulaciones.

6 - Dos formas de ver el p-valor

a)

Queremos hallar $P\ (T \geq T_{obs} \mid H_0\ es\ Verdadera)$.

$$T = \frac{\overline{X} - \mu_0}{\sqrt{\frac{\sigma_0^2}{n}}} \sim N(0,1)$$

En el ejercicio 4-a obtuvimos $\overline{X}_{obs} = 42,9895 \Rightarrow T_{obs} = \frac{42,9895 - 41,98465}{0,5120183} = 1,962528$.

Entonces, $P\ (T \geq 1,962528) = 0,02485052$ (Bajo H_0). Este número representa el p-valor de la muestra nueva.

b)

El valor obtenido en el ejercicio 6-a representa el menor nivel de significancia α para el que rechazamos H_0 con los datos observados. En la tabla del ejercicio 4-d se puede ver que cuando el nivel de significancia α es menor que el p-valor, se toma la decisión de no rechazar la hipótesis nula, y cuando es mayor, se rechaza H_0.

7- ¿Cómo sé que los datos son normales?

a)

Sea $X_i = "'play - delay' \text{ de un usuario aleatorio de la nueva versión}", 1 \leq i \leq 200,$

Quiero un test de nivel de significancia α para las hipótesis

$$H_o: X \sim N(\mu, \sigma_0^2) \text{ vs. } H_1: X \text{ no tiene distribución } N(\mu, \sigma_0^2)$$

Sea $X_i = "'play - delay' \text{ de un usuario aleatorio de la versión histórica}",$ $1 \leq i \leq 3498,$

Quiero un test de nivel de significancia α para las hipótesis

$$H_o: X \sim N(\mu_0, \sigma_0^2) \text{ vs. } H_1: X \text{ no tiene distribución } N(\mu_0, \sigma_0^2)$$

b) Realizando el test en R, obtenemos un p-valor = 0,6934 para los datos históricos y p-valor = 0,867. En ambos casos, obtenemos un p-valor por encima de 0,5, por lo que podemos decir que no hay evidencia suficiente para rechazar H_0 (no rechazo H_0a nivel 0,05).