Unidad 6: Inferencia Estadística: Pruebas de hipótesis paramétricas

Mg. J. Eduardo Gamboa U.

Table of contents

Descripción del caso

Mediante el archivo `datos_u6.csv` se busca analizar el consumo mensual de energía junto a diversas características de un edificio y factores ambientales. Contiene datos de varios tipos de edificios, el metraje cuadrado, el número de ocupantes, los electrodomésticos utilizados, la temperatura promedio y el día de la semana.

Carga de paquetes y lectura de datos

```
library(pacman)
p_load(dplyr, ggplot2, magrittr, EnvStats, binom)

datos <- read.csv('datos_u6.csv')
datos |> rename("TipoEdificio" = 1, "AreaPies2" = 2, "Ocupantes" = 3, "Electro" = 4,
                "Temperat" = 5, "DiaSemana" = 6, "Consumo" = 7) -> datos
```

Prueba de hipótesis para una media

¿El consumo promedio mensual de energía eléctrica es superior a los 4000 Kwh?

$$H_0 : \mu \leq 4000 \qquad H_1 : \mu > 4000 \qquad \alpha = 0.05$$

```
datos |>
  pull(Consumo) |>
  t.test(alternative = "greater", mu = 4000)
```

```

	One Sample t-test

data:  pull(datos, Consumo)
t = 5.633, df = 999, p-value = 1.151e-08
alternative hypothesis: true mean is greater than 4000
95 percent confidence interval:
 4117.661      Inf
sample estimates:
mean of x
 4166.253
```

El pvalor se obtiene de la siguiente manera:

$$pv = P(t_{999} > 5.633) = 1.151 \times 10^{-8}$$

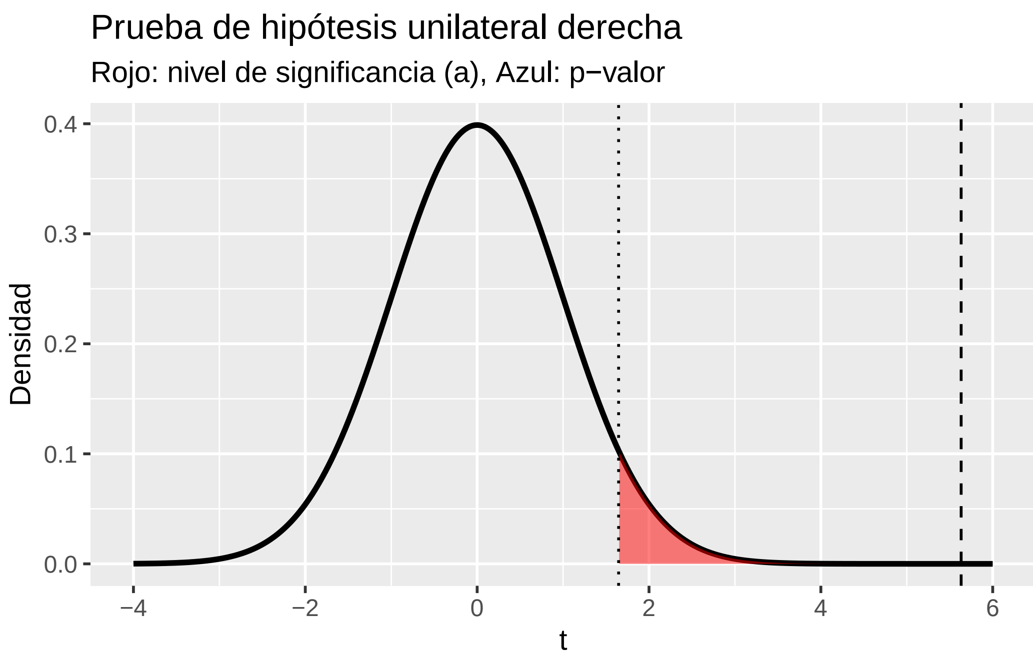
```
pt(q = 5.633, df = 999,lower.tail = F)
```

```
[1] 1.150953e-08
```

```r
t_obs <- 5.633
alpha <- 0.05

x_vals   <- seq(-4, 6, length.out = 1000)
dens_df  <- data.frame(x = x_vals, y = dt(x_vals, 999))
alpha_df <- subset(dens_df, x >= qt(1 - alpha, 999))
pval_df  <- subset(dens_df, x >= t_obs)

ggplot(dens_df, aes(x, y)) +
  geom_line(linewidth = 1) +
  geom_area(data = alpha_df, aes(x, y), fill = "red", alpha = 0.5) +
  geom_area(data = pval_df, aes(x, y), fill = "blue", alpha = 0.5) +
  geom_vline(xintercept = t_obs, linetype = "dashed") +
  geom_vline(xintercept = qt(1 - alpha, 999), linetype = "dotted") +
  labs(
    title = "Prueba de hipótesis unilateral derecha",
    subtitle = "Rojo: nivel de significancia ( ), Azul: p-valor",
    x = "t", y = "Densidad")
```



Decisión: Se rechaza H_0

Conclusión: Con un nivel de significancia del 5%, sí existe suficiente evidencia estadística para afirmar que el consumo promedio mensual de energía eléctrica es superior a los 4000 Kwh

Prueba de hipótesis para una varianza

¿La variabilidad de las temperaturas es menor a 50 °C^2?

$$H_0 : \sigma^2 \geq 50 \qquad H_1 : \sigma^2 < 50 \qquad \alpha = 0.05$$

```r
library(DescTools)
datos |>
  pull(Temperat) |>
  varTest(sigma.squared = 50, alternative = "less")
```

```
$statistic
Chi-Squared
   1018.556

$parameters
 df
999

$p.value
[1] 0.6734908

$estimate
variance
50.97878

$null.value
variance
      50

$alternative
[1] "less"

$method
[1] "Chi-Squared Test on Variance"

$data.name
[1] "pull(datos, Temperat)"

$conf.int
     LCL      UCL
 0.00000 54.96016
```

```
attr(,"conf.level")
[1] 0.95

attr(,"class")
[1] "htestEnvStats"
```

El pvalor se obtiene de la siguiente manera: $pv = P(\chi^2_{999} < 1018.556)$

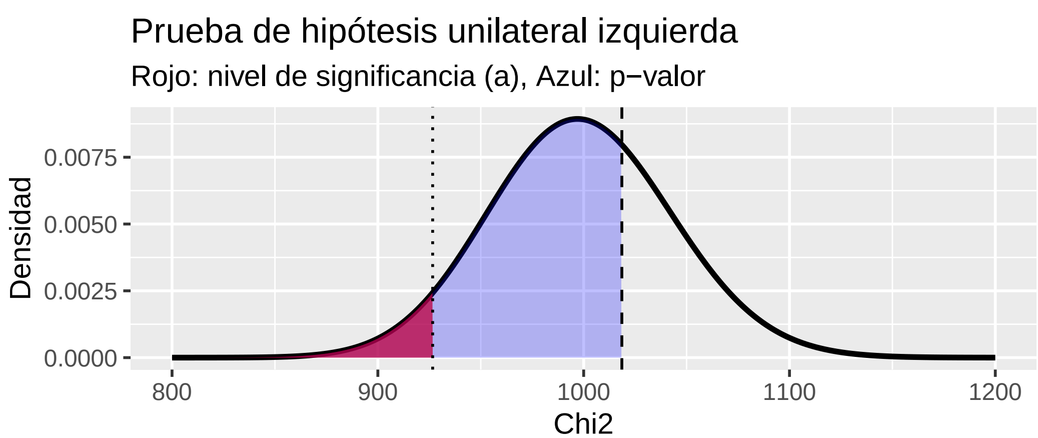
```r
pchisq(q = 1018.556, df = 999)
```

```
[1] 0.6734907
```

```r
chi_obs <- 1018.556; alpha   <- 0.05

x_vals   <- seq(800, 1200, length.out = 1000)
dens_df  <- data.frame(x = x_vals, y = dchisq(x_vals, 999))
alpha_df <- subset(dens_df, x <= qchisq(alpha, 999))
pval_df  <- subset(dens_df, x <= chi_obs)

ggplot(dens_df, aes(x, y)) +
  geom_line(linewidth = 1) +
  geom_area(data = alpha_df, aes(x, y), fill = "red", alpha = 0.75) +
  geom_area(data = pval_df, aes(x, y), fill = "blue", alpha = 0.25) +
  geom_vline(xintercept = chi_obs, linetype = "dashed") +
  geom_vline(xintercept = qchisq(alpha, 999), linetype = "dotted") +
  labs(title = "Prueba de hipótesis unilateral izquierda",
       subtitle = "Rojo: nivel de significancia ( ), Azul: p-valor",
       x = "Chi2", y = "Densidad")
```



Decisión: No se rechaza H_0

Conclusión: Con un nivel de significancia del 5%, no existe suficiente evidencia estadística para afirmar que la variabilidad de las temperaturas es menor a 50 $°C^2$

Prueba de hipótesis para una proporción

Se dice que si un inmueble posee un área superior a los 15 mil pies cuadrados se le considera "grande". ¿Se puede señalar que la proporción de inmuebles grandes es del 60%?

$$H_0 : \pi = 0.60 \qquad H_1 : \pi \neq 0.60 \qquad \alpha = 0.05$$

```
datos |>
  mutate(Tamano = ifelse(AreaPies2>15000, "Grande", "No grande")) |>
  count(Tamano) |> mutate(Prop = n/sum(n)) |> pull(n) -> xn

prop.test(x = xn[1], n = xn[1] + xn[2], p = 0.60, alternative = "two.sided", correct = F)
```

```
    1-sample proportions test without continuity correction

data:  xn[1] out of xn[1] + xn[2], null probability 0.6
X-squared = 51.337, df = 1, p-value = 7.778e-13
alternative hypothesis: true p is not equal to 0.6
95 percent confidence interval:
 0.6821396 0.7382455
sample estimates:
    p
0.711
```

El pvalor se obtiene de la siguiente manera:

$$Z_{calc} = \frac{0.711 - 0.6}{\sqrt{\frac{0.6 \times 0.4}{1000}}} = 7.165019$$

$$pv = 2P(Z > 7.165019) = 7.778 \times 10^{-13}$$

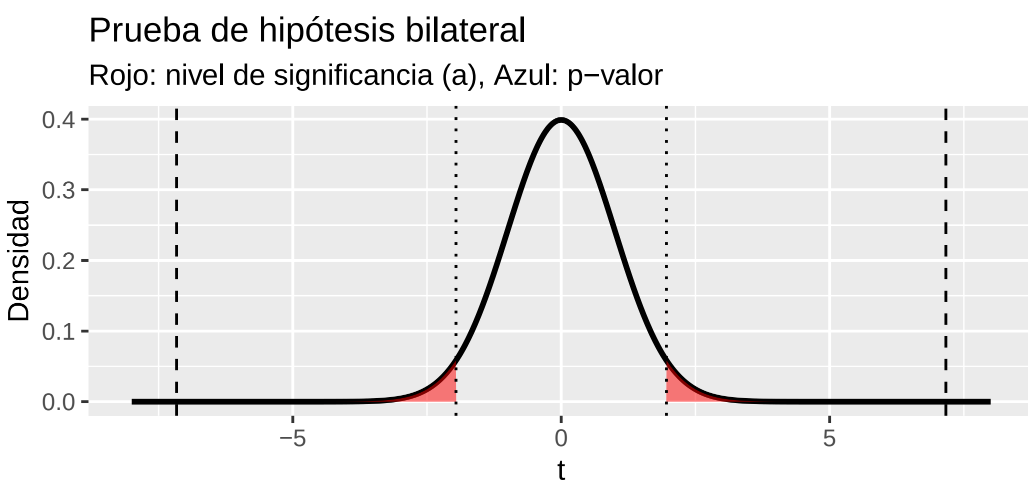
```
2*pnorm(q = 7.165019, lower.tail = F)
```

```
[1] 7.777577e-13
```

```r
z_obs <- 7.165019
alpha <- 0.05

x_vals   <- seq(-8, 8, length.out = 1000)
dens_df  <- data.frame(x = x_vals, y = dnorm(x_vals))
alpha_df_left <- subset(dens_df, x <= qnorm(alpha / 2))
alpha_df_right <- subset(dens_df, x >= qnorm(1 - alpha / 2))
pval_df_left <- subset(dens_df, x <= -abs(z_obs))
pval_df_right <- subset(dens_df, x >= abs(z_obs))

ggplot(dens_df, aes(x, y)) +
  geom_line(linewidth = 1) +
 geom_area(data = alpha_df_left, aes(x, y), fill = "red", alpha = 0.5) +
  geom_area(data = alpha_df_right, aes(x, y), fill = "red", alpha = 0.5) +
  geom_area(data = pval_df_left, aes(x, y), fill = "blue", alpha = 0.5) +
  geom_area(data = pval_df_right, aes(x, y), fill = "blue", alpha = 0.5) +
  geom_vline(xintercept = z_obs, linetype = "dashed") +
  geom_vline(xintercept = -z_obs, linetype = "dashed") +
  geom_vline(xintercept = qnorm(1 - alpha/2), linetype = "dotted") +
  geom_vline(xintercept = qnorm(alpha/2), linetype = "dotted") +
  labs(title = "Prueba de hipótesis bilateral",
       subtitle = "Rojo: nivel de significancia ( ), Azul: p-valor",
       x = "t", y = "Densidad")
```



Decisión: Se rechaza H_0

Conclusión: Con un nivel de significancia del 5%, no existe suficiente evidencia estadística para afirmar que la proporción de inmuebles grandes es del 60%.

Prueba de hipótesis para dos medias independientes

¿El consumo medio de energía en los edificios industriales supera en más de 500 Kwh a los comerciales? Considerar $\alpha = 0.01$.

Primero, se prueba la igualdad de varianzas:

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1 \qquad H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1 \qquad \alpha = 0.01$$

```
datos |> filter(TipoEdificio %in% c("Commercial","Industrial")) -> datos_2medias
var.test(formula     = Consumo ~ TipoEdificio,
         data        = datos_2medias,
         alternative = "two.sided")
```

```

	F test to compare two variances

data:  Consumo by TipoEdificio
F = 0.95337, num df = 335, denom df = 316, p-value = 0.6662
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.7665537 1.1848793
sample estimates:
ratio of variances
          0.953368
```

```
# Notar que en vez de declarar las dos muestras, estamos indicando
# variable y ~ variable de grupos
```

¿Cómo se calculó el p-valor?

$$pv = 2P(F_{335,316} < 0.95337)$$

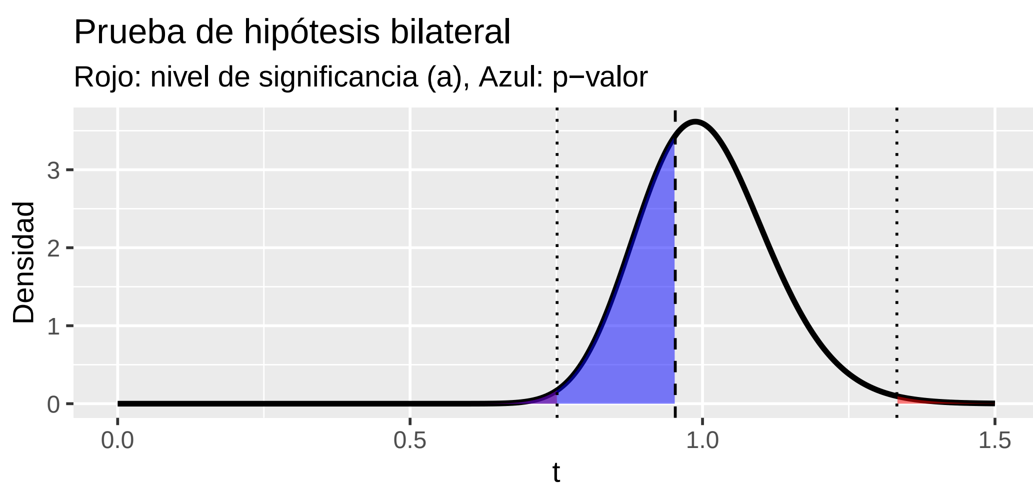
```
2*pf(0.95337,335,316)
```

```
[1] 0.6661716
```

```r
f_obs <- 0.95337
alpha <- 0.01

x_vals   <- seq(0,1.5, length.out = 1000)
dens_df  <- data.frame(x = x_vals, y = df(x_vals,335,316))
alpha_df_left <- subset(dens_df, x <= qf(alpha / 2, 335, 316))
alpha_df_right <- subset(dens_df, x >= qf(1 - alpha / 2, 335, 316))
pval_df <- subset(dens_df, x <= f_obs)

ggplot(dens_df, aes(x, y)) +
  geom_line(linewidth = 1) +
  geom_area(data = alpha_df_left, aes(x, y), fill = "red", alpha = 0.5) +
  geom_area(data = alpha_df_right, aes(x, y), fill = "red", alpha = 0.5) +
  geom_area(data = pval_df, aes(x, y), fill = "blue", alpha = 0.5) +
  geom_vline(xintercept = f_obs, linetype = "dashed") +
  geom_vline(xintercept = qf(1 - alpha/2, 335, 316), linetype = "dotted") +
  geom_vline(xintercept = qf(alpha/2, 335, 316), linetype = "dotted") +
  labs(title = "Prueba de hipótesis bilateral",
       subtitle = "Rojo: nivel de significancia ( ), Azul: p-valor",
       x = "t", y = "Densidad")
```



Dado que $pv > \alpha$, entonces no se rechaza H_0. En conclusión, las varianzas son homogéneas. Sabiendo esto, se pasa a comparar las medias.

El enunciado señala que se debe verificar la afirmación $\mu_I - \mu_C > 500$, sin embargo R toma por defecto los niveles en orden alfabético, es decir primero "Commercial" y luego "Industrial". Hay 2 opciones: Reordenar los niveles, o reordenar la hipótesis. Eligiendo esta última opción tendríamos que la afirmación inicial es equivalente a $\mu_C - \mu_I < -500$.

Por lo tanto:

$$H_0 : \mu_C - \mu_I \geq -500 \qquad H_1 : \mu_C - \mu_I < -500 \qquad \alpha = 0.01$$

```
t.test(formula = Consumo ~ TipoEdificio, data = datos_2medias, mu = -500,
       alternative = "less", var.equal = T)
```

```
    Two Sample t-test

data:  Consumo by TipoEdificio
t = -1.6384, df = 651, p-value = 0.05091
alternative hypothesis: true difference in means between group Commercial and group Industria
95 percent confidence interval:
      -Inf -499.4383
sample estimates:
mean in group Commercial mean in group Industrial
                4130.024                 4735.143
```

¿Cómo se calculó el p-valor?

$$pv = 2P(t_{651} < -1.6384) = 0.0509$$

```
pt(q = -1.6384, df = 651)
```

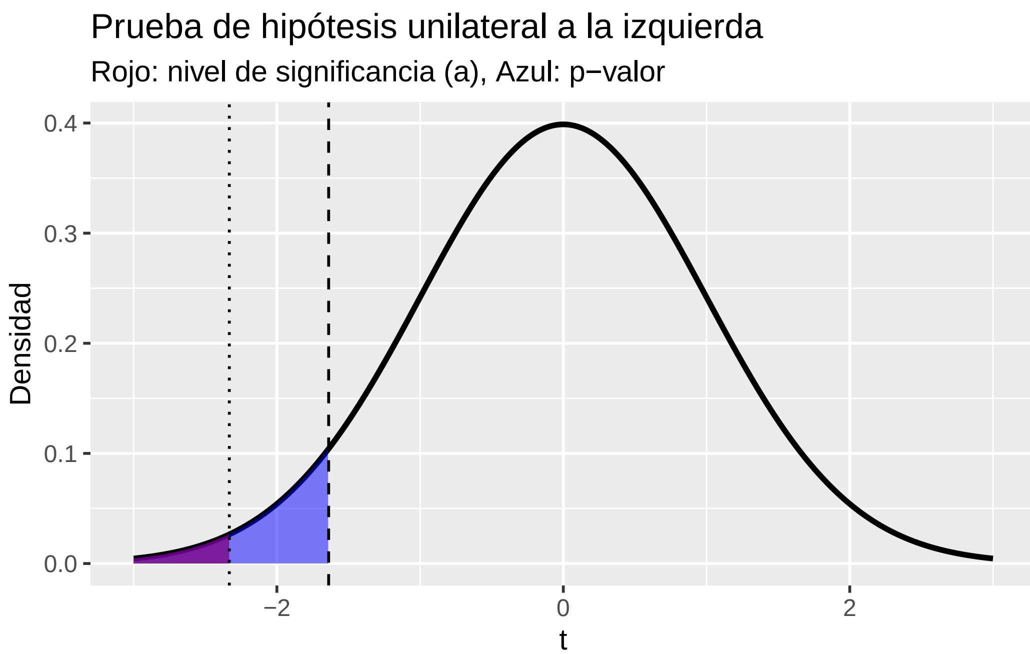
```
[1] 0.05091073
```

```
t_obs <- -1.6384
alpha <- 0.01

x_vals   <- seq(-3,3, length.out = 1000)
dens_df  <- data.frame(x = x_vals, y = dt(x_vals,651))
alpha_df <- subset(dens_df, x <= qt(alpha, 651))
pval_df  <- subset(dens_df, x <= t_obs)

ggplot(dens_df, aes(x, y)) +
  geom_line(linewidth = 1) +
  geom_area(data = alpha_df, aes(x, y), fill = "red", alpha = 0.75) +
  geom_area(data = pval_df, aes(x, y), fill = "blue", alpha = 0.5) +
  geom_vline(xintercept = t_obs, linetype = "dashed") +
```

```r
  geom_vline(xintercept = qt(alpha, 651), linetype = "dotted") +
  labs(title = "Prueba de hipótesis unilateral a la izquierda",
       subtitle = "Rojo: nivel de significancia ( ), Azul: p-valor",
       x = "t", y = "Densidad")
```



Decisión: Dado que $pv > \alpha$, no se rechaza la hipótesis nula.

Conclusión: el consumo medio de energía en los edificios industriales no supera en más de 500 Kwh a los comerciales.

Prueba de hipótesis para dos proporciones

Se dice que si un inmueble posee un área superior a los 15 mil pies cuadrados se le considera "grande". ¿Se puede señalar que la proporción de inmuebles comerciales grandes es mayor que los industriales?

$$H_0 : \pi_1 - \pi_2 \leq 0 \qquad H_1 : \pi_1 - \pi_2 > 0 \qquad \alpha = 0.05$$

```r
datos |>
  filter(TipoEdificio %in% c("Commercial","Industrial")) |>
  mutate(Tamano = ifelse(AreaPies2>15000, "Grande", "No grande")) |>
  group_by(TipoEdificio) |>
  count(Tamano) |> mutate(Prop = n/sum(n)) -> tabla

tabla
```

```
# A tibble: 4 x 4
# Groups:   TipoEdificio [2]
  TipoEdificio Tamano        n  Prop
  <chr>        <chr>     <int> <dbl>
1 Commercial   Grande      233 0.693
2 Commercial   No grande   103 0.307
3 Industrial   Grande      231 0.729
4 Industrial   No grande    86 0.271
```

```r
prop.test(x = tabla$n[c(1,3)],
          n = tabla$n[c(1,3)]+tabla$n[c(2,4)],
          alternative = "greater", correct = F)
```

```

    2-sample test for equality of proportions without continuity correction

data:  tabla$n[c(1, 3)] out of tabla$n[c(1, 3)] + tabla$n[c(2, 4)]
X-squared = 0.98572, df = 1, p-value = 0.8396
alternative hypothesis: greater
95 percent confidence interval:
 -0.09355509  1.00000000
sample estimates:
   prop 1    prop 2
0.6934524 0.7287066
```

El pvalor se halla de la siguiente manera:

$$P(Z > -0.9928357) = P(\chi^2_1 > 0.98572) = 0.8396$$

```r
p1 = 233/336 # 0.69345
p2 = 231/317 # 0.72871
p  = 464/653 # 0.71057
(zc = (p1-p2)/sqrt(p*(1-p)*(1/336+1/317)))
```

```
[1] -0.9928357
```

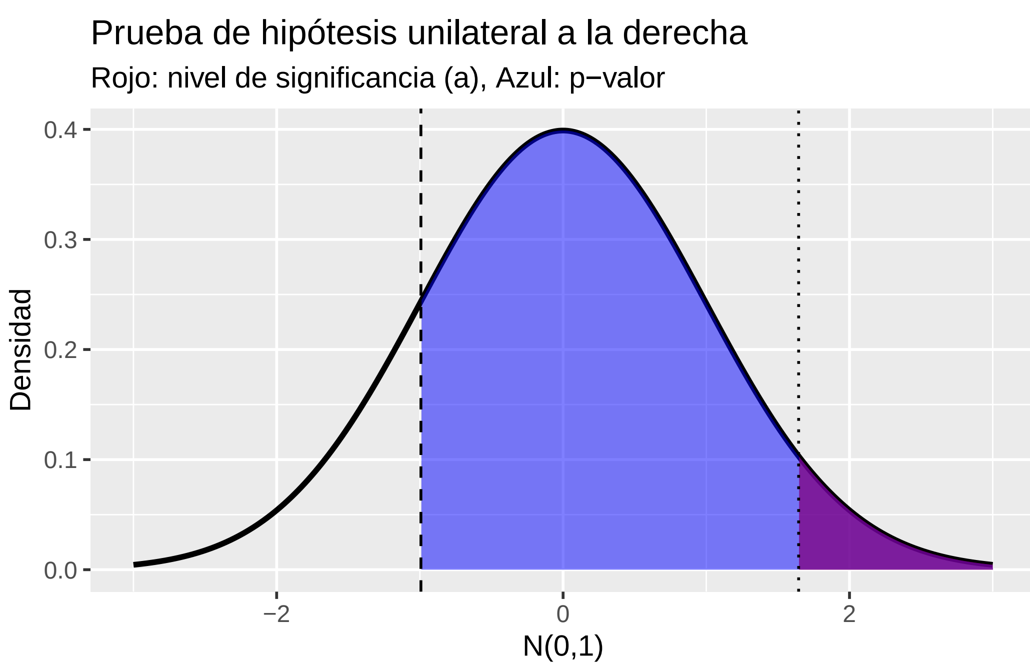
```r
pnorm(q = zc, lower.tail = F)
```

```
[1] 0.839605
```

```r
z_obs <- -0.9928357
alpha <- 0.05

x_vals   <- seq(-3,3, length.out = 1000)
dens_df  <- data.frame(x = x_vals, y = dnorm(x_vals))
alpha_df <- subset(dens_df, x >= qnorm(1-alpha))
pval_df  <- subset(dens_df, x >= z_obs)

ggplot(dens_df, aes(x, y)) +
  geom_line(linewidth = 1) +
  geom_area(data = alpha_df, aes(x, y), fill = "red", alpha = 0.75) +
  geom_area(data = pval_df, aes(x, y), fill = "blue", alpha = 0.5) +
  geom_vline(xintercept = z_obs, linetype = "dashed") +
  geom_vline(xintercept = qnorm(1-alpha), linetype = "dotted") +
  labs(title = "Prueba de hipótesis unilateral a la derecha",
       subtitle = "Rojo: nivel de significancia ( ), Azul: p-valor",
       x = "N(0,1)", y = "Densidad")
```



Decisión: Dado que $pv > \alpha$, no se rechaza la hipótesis nula.

Conclusión: No existe evidencia estadística para señalar que la proporción de inmuebles comerciales grandes es mayor que los industriales