XS3310 Teoría Estadística

I Semestre 2021

Escuela de Estadística

2021-05-24

class: center, middle

¿Qué hemos visto hasta ahora?

Todo sobre estimadores puntuales + pivotes e intervalos de confianza. Bootstrap y una introducción a los contrastes de hipótesis.

¿Qué vamos a discutir hoy?

Contrastes de hipótesis: función de potencia, tamaño del contraste, el valor p.

Los datos de la cantidad de lluvia para este experimento están acá.

```
nubes <- read.table(file = "./data/clouds.txt", sep = "\t",header = TRUE)
head(nubes)
```

```
##   Unseeded.Clouds Seeded.Clouds
## 1          1202.6        2745.6
## 2           830.1        1697.8
## 3           372.4        1656.0
## 4           345.5         978.0
## 5           321.2         703.4
## 6           244.3         489.1
```

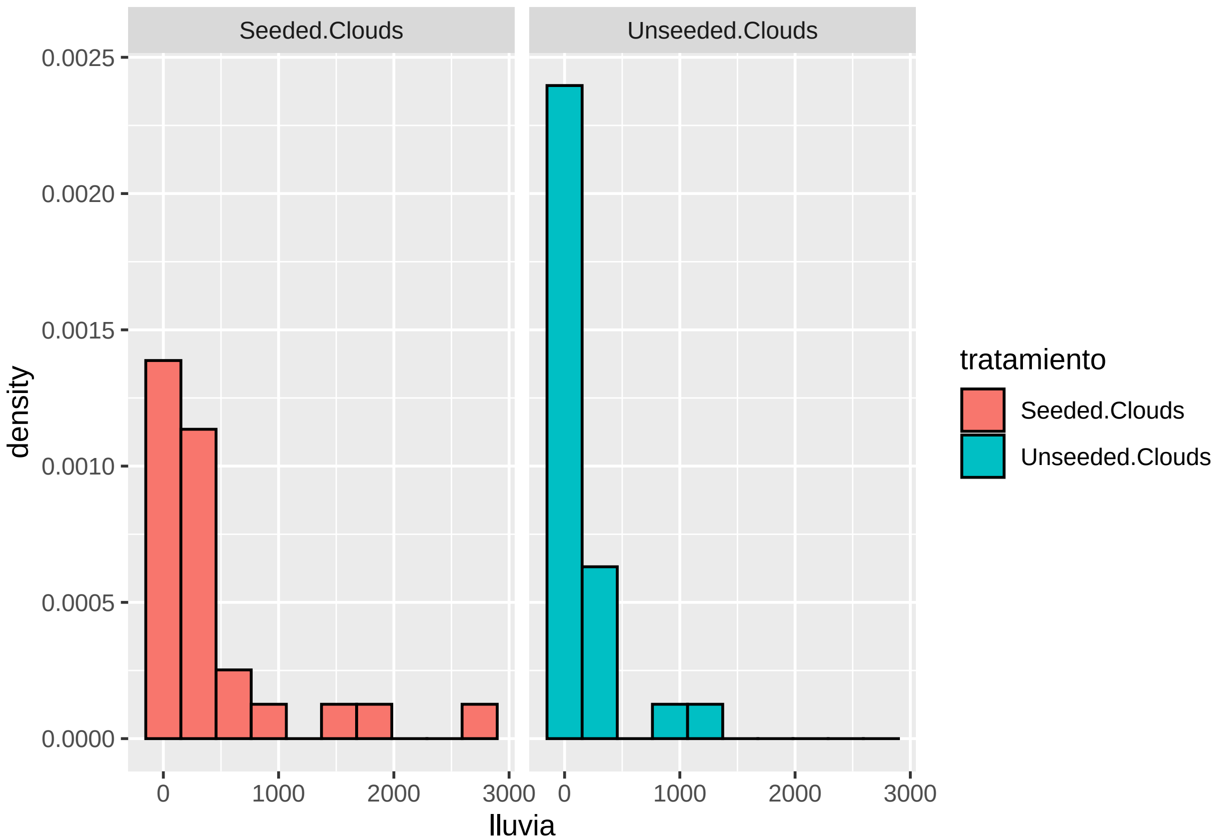
Sin embargo usaremos los datos en escala logarítmica para facilitar el cálculo

```
lognubes <- log(nubes)
head(lognubes)
```

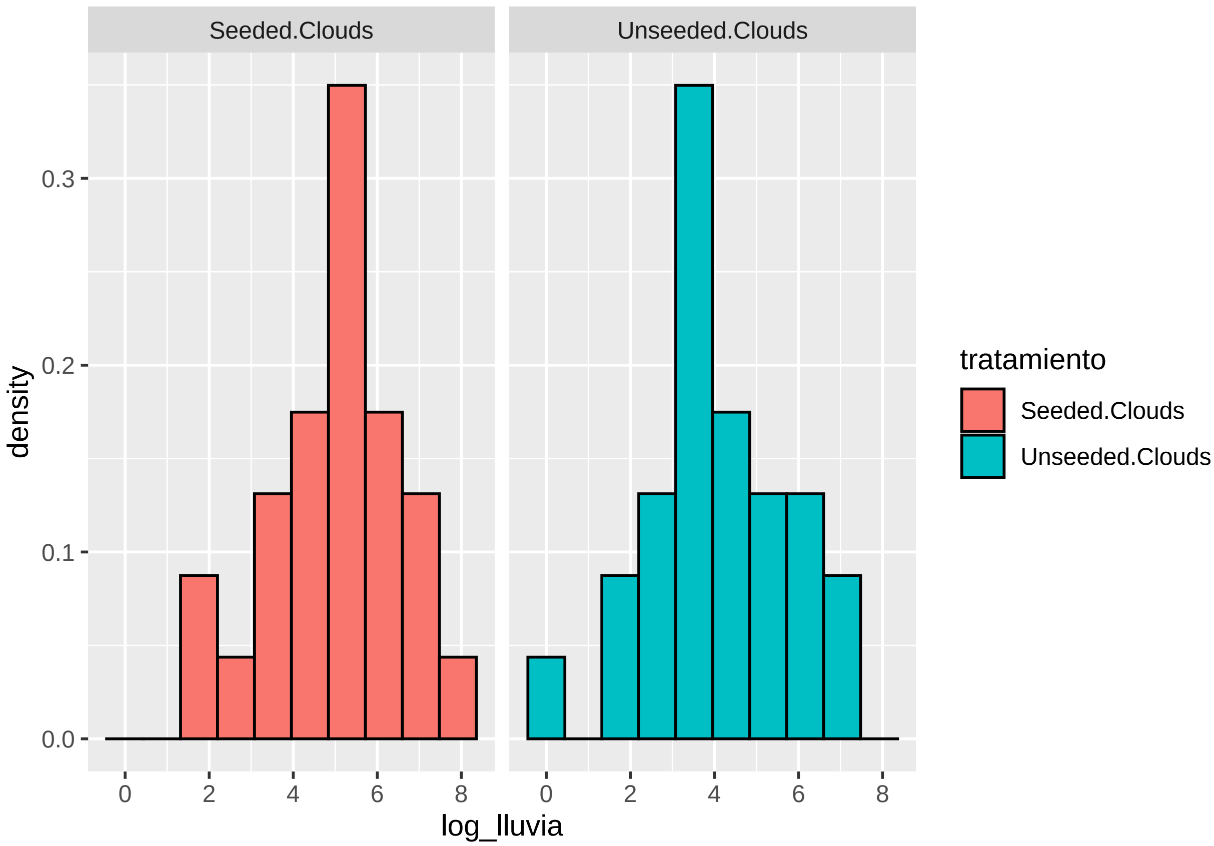
```
##   Unseeded.Clouds Seeded.Clouds
## 1        7.092241      7.917755
## 2        6.721546      7.437089
## 3        5.919969      7.412160
## 4        5.844993      6.885510
## 5        5.772064      6.555926
## 6        5.498397      6.192567
```

Observe que el comportamiento es distinto en ambos casos.

```
df <- as.data.frame(nubes) %>%
    pivot_longer(
        cols = everything(),
        names_to = "tratamiento", values_to = "lluvia"
    ) %>%
    mutate(log_lluvia = log(lluvia))
ggplot(data = df) +
    geom_histogram(aes(
        x = lluvia,
        y = ..density..,
        fill = tratamiento
    ),
    color = "black", bins = 10
    ) +
    facet_wrap(. ~ tratamiento)
```




```
ggplot(data = df) +
    geom_histogram(aes(
        x = log_lluvia,
        y = ..density..,
        fill = tratamiento
    ), color = "black", bins = 10) +
    facet_wrap(. ~ tratamiento)
```



En este caso supondremos que la variable `log_lluvia` se puede modelar como una $N(\mu, \sigma^2)$, μ, σ desconocidos.

Preguntamos lo siguiente:

En este caso sería $\theta = (\mu, \sigma^2)$, ¿Será cierto que para $\theta \in \{(\mu, \sigma^2) : \mu > 4\}$?

El valor de $\mu > 4$ nace a partir de una pregunta de investigación y se fórmula una **hipótesis** con respecto a los datos.

En este caso el valor donde decrece la curva es cercano a 0. Eso quiere decir que antes de ese valor, nos encontramos en la región de rechazo. Luego esa región se va haciendo cada vez más pequeña $|\overline{X} - \mu| \approx 0$.

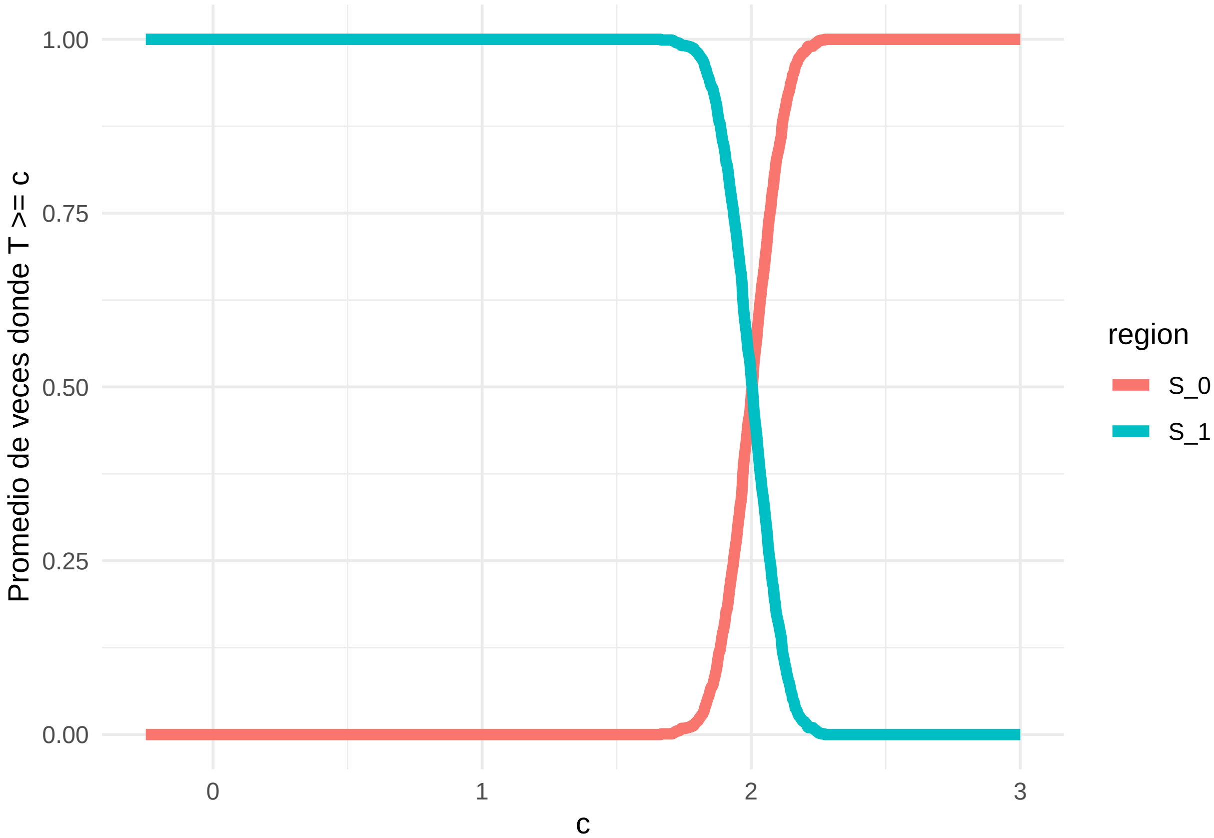
Ojo lo que pasaría si por ejemplo cambiamos a $\mu = 4$,

```
X <- matrix(rnorm(1000 * 1000, mean = 2, sd = 3), ncol = 1000)

Xbar <- apply(X, 2, mean)

mu0 <- 4
T <- abs(Xbar - mu0)
c <- seq(-0.25, 3, length.out = 1000)
df <- data.frame(c = numeric(), test = logical(), region = character())
for (k in 1:length(c)) {
    df <- rbind(
        df,
        data.frame(c = c[k], test = mean(T >= c[k]), region = "S_1")
    )
}
df <- rbind(df, data.frame(c, test = 1 - df$test, region = "S_0"))
```

```r
ggplot(df, aes(x = c, y = test, color = region)) +
    geom_line(size = 2) +
    ylab("Promedio de veces donde T >= c") +
    theme_minimal()
```



El valor donde comienza a crecer la curva se desvía a un valor cercano a 2.

Para el caso del ejemplo de la lluvia definimos que

$$H_0 : \mu \leq 4 \text{ versus } H_1 : \mu > 4$$

En este caso podríamos decir que rechazamos H_0 si la media empírica es "más grande" que 4 y no rechazamos H_0 si la media empírica es "más pequeña" que 4.

El problema acá es que "más grande" y "más pequeña" no son términos precisos.

Tenemos dos opciones

Construya la región de critica de la forma

$$S_0 = \left\{ \boldsymbol{x} :\leq \bar{X}_n - \mu_0 \leq c \right\}, \quad \text{y} \quad S_1 = S_0^C \tag{1}$$

y observe cuál es la probabilidad que ocurra para cada tipo de c. El problema con esta construcción es que requiere conocer todos los posibles vectores de datos $\mathbf{X}$ y construir los conjuntos S_0 y S_1.

Una mejor opción es tener un estadístico sencillo que cumpla dos condiciones:

1. Un estadístico sencillo de calcular posiblemente suficiente, minimal y eficiente.

2. Un estadístico con una distribución conocida.

En ese caso $\overline{X}_n$ funciona muy bien, porque tiene todas las buenas propiedades de sufiencia, minimalidad y eficiencia, y además sabemos su distribución según lo estudiando en capítulos pasados. Entonces

$$U = \frac{n^{1/2}(\overline{X}_n - \mu_0)}{\sigma'} \sim t_{n-1}$$

Lo natural debería ser rechazar H_0 si U es grande.

```r
colnames(lognubes)
```

```
## [1] "Unseeded.Clouds" "Seeded.Clouds"
```

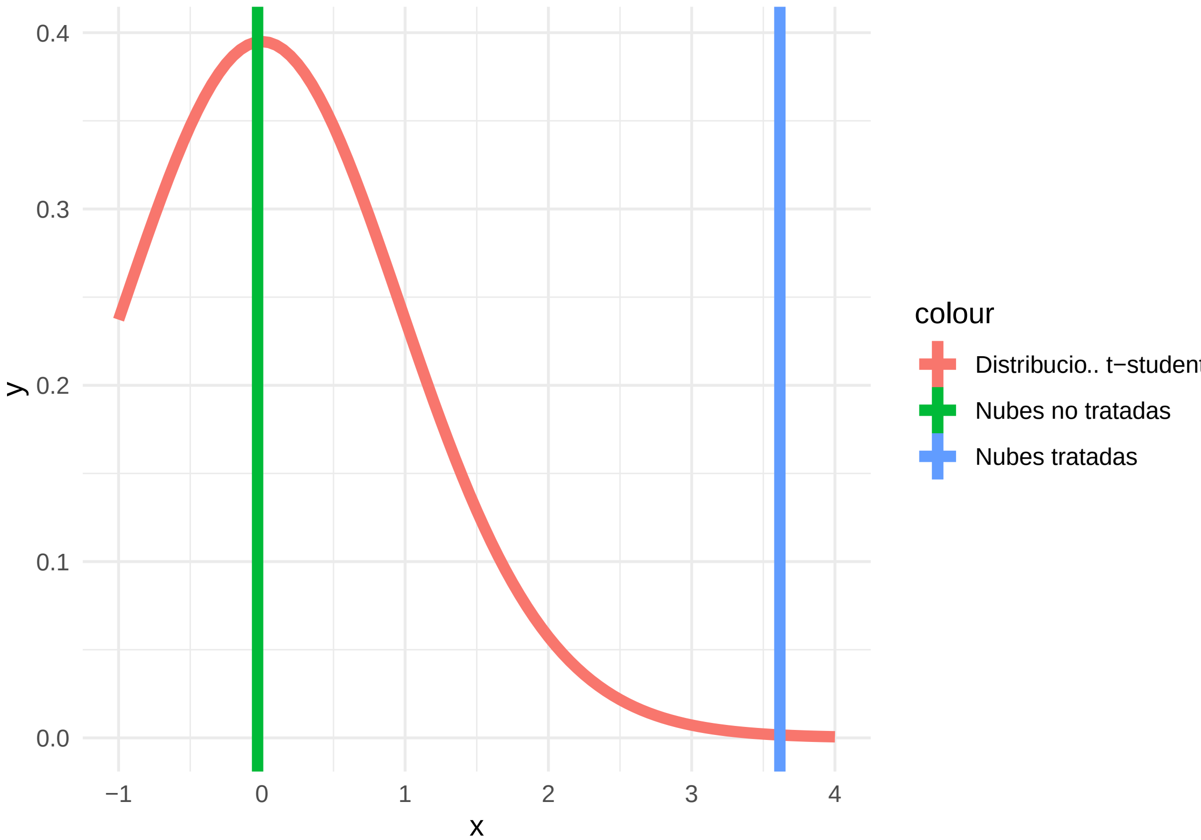
```r
Xbarra1 <- mean(lognubes[, 1])
Xbarra2 <- mean(lognubes[, 2])
sigma_prima1 <- sd(lognubes[, 1])
sigma_prima2 <- sd(lognubes[, 2])
n <- dim(lognubes)[1]
(U1 <- sqrt(n) * (Xbarra1 - 4) / sigma_prima1)
```

```
## [1] -0.02979683
```

```r
(U2 <- sqrt(n) * (Xbarra2 - 4) / sigma_prima2)
```

```
## [1] 3.615624
```

```r
ggplot(data = data.frame(x = (c(-1, 4))), mapping = aes(x)) +
    stat_function(
        fun = dt, args = list(df = n - 1),
        mapping = aes(color = "Distribucioń t-student"), size = 2
    ) +
    geom_vline(mapping = aes(
        xintercept = U1,
        color = "Nubes no tratadas"
    ), size = 2) +
    geom_vline(mapping = aes(
        xintercept = U2,
        color = "Nubes tratadas"
    ), size = 2) +
    theme_minimal()
```



- Estadístico de prueba: $T = |\bar{X}_n - \mu_0|$.
- Región de rechazo: $R = (c, \infty)$.

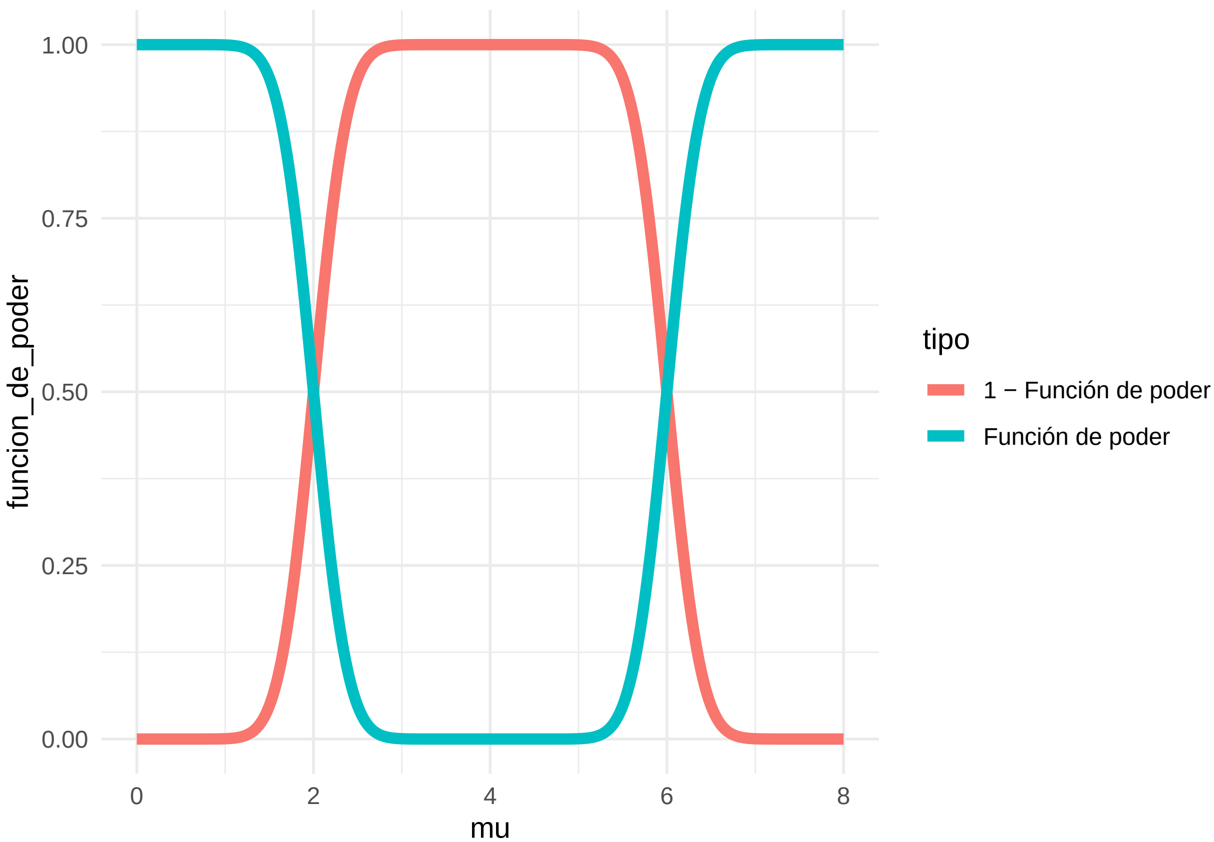
Como $X_1, \dots, X_n \sim N(\mu, \sigma^2)$, μ desconocido, σ^2 conocido entonces $\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

- Función de potencia:

$$
\begin{aligned}
\pi(\theta|\delta) = \mathbb{P}[T \in R|\mu] &= \mathbb{P}[|\bar{X}_n - \mu_0| > c|\mu] \\
&= \mathbb{P}[\bar{X}_n > \mu_0 + c|\mu] + \mathbb{P}[\bar{X}_n < \mu_0 - c|\mu] \\
&= \mathbb{P}\left[\sqrt{n}\frac{(\bar{X}_n - \mu)}{\sigma} > \frac{(\mu_0 + c - \mu)}{\sigma}\sqrt{n}\middle|\mu\right] + \mathbb{P}\left[\sqrt{n}\frac{(\bar{X}_n - \mu)}{\sigma} < \frac{(\mu_0 - c - \mu)}{\sigma}\sqrt{n}\middle|\mu\right] \\
&= 1 - \Phi\left(\sqrt{n}\frac{(\mu_0 + c - \mu)}{\sigma}\right) + \Phi\left(\sqrt{n}\frac{(\mu_0 - c - \mu)}{\sigma}\right)
\end{aligned}
$$

```
mu0 <- 4
c <- 2
n <- 100
sigma <- 3
mu <- seq(0, 8, length.out = 1000)
funcion_de_poder <- 1 -
    pnorm(sqrt(n) * (mu0 + c - mu) / sigma) +
```

```r
    pnorm(sqrt(n) * (mu0 - c - mu) / sigma)
df <- data.frame(mu, funcion_de_poder, tipo = "Función de poder")
df <- rbind(df, data.frame(mu,
    funcion_de_poder = 1 - df$funcion_de_poder,
    tipo = "1 - Función de poder"
))
ggplot(df, aes(mu, funcion_de_poder, color = tipo)) +
    geom_line(size = 2) +
    theme_minimal()
```




```r
mu <- seq(0, 8, length.out = 100)
c <- seq(0, 4, length.out = 100)
mu_c <- expand.grid(mu, c)
funcion_de_poder_n_c <- 1 -
    pnorm(sqrt(n) * (mu0 + mu_c[, 2] - mu_c[, 1]) / sigma) +
    pnorm(sqrt(n) * (mu0 - mu_c[, 2] - mu_c[, 1]) / sigma)
```

Tipos de error:

En este tipo de pruebas se puede cometer dos tipos de errores,

- *Error Tipo I*: error de rechazar H_0 si $\theta \in \Omega_0$.
- *Error Tipo II*: error de no rechazar H_0 si $\theta \in \Omega_1$ en términos de la función de potencia.

En términos de la función de poder tenemos que

- Si $\theta \in \Omega_0$: $\pi(\theta|\delta)$ es el error tipo I.
- Si $\theta \in \Omega_1$: $1 - \pi(\theta|\delta)$ es el error tipo II.

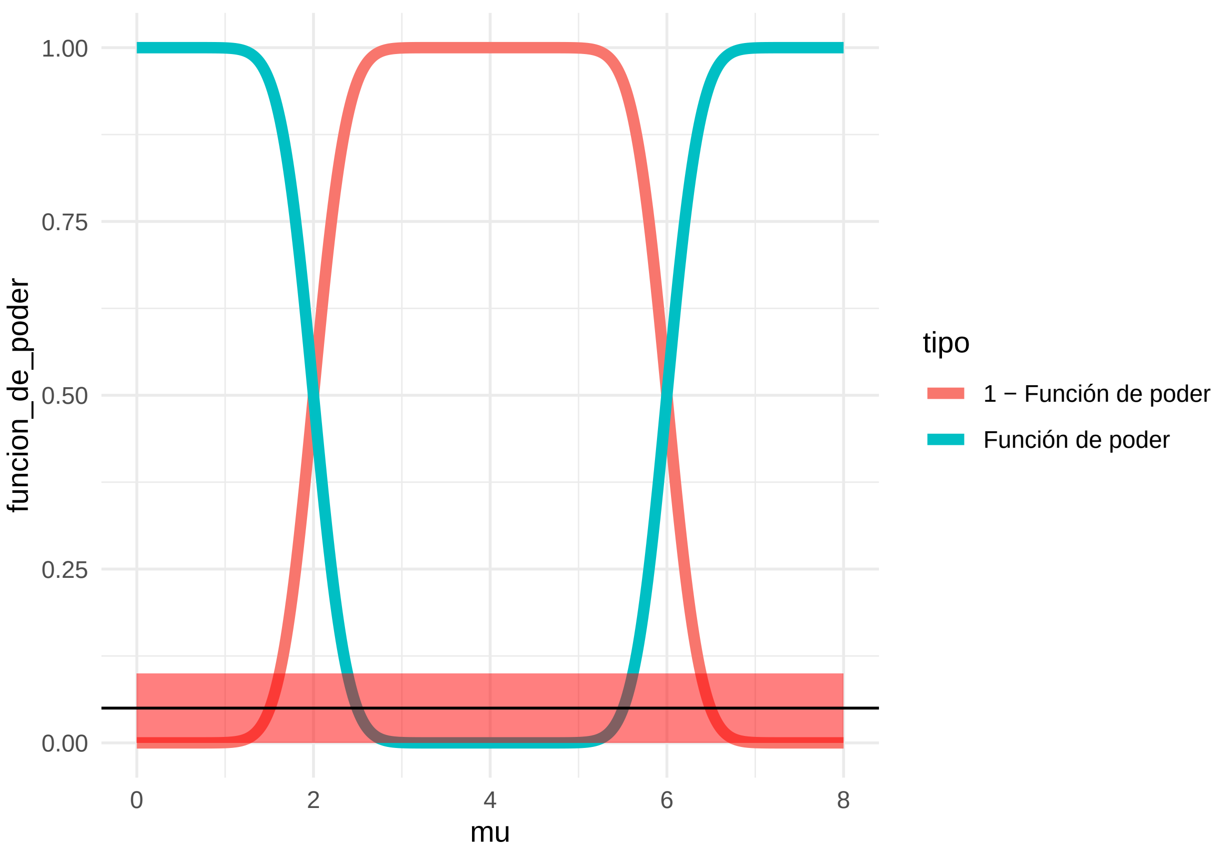
El objetivo es hacer $\pi(\theta|\delta)$ pequeño cuando $\theta \in \Omega_0$. También se requiere que $\pi(\theta|\delta)$ sea grande cuando $\theta \in \Omega_1$. Una forma de alcanzar ese balance es seleccionar $\alpha_0 \in (0, 1)$ tal que

$$\pi(\theta|\delta) \leq \alpha_0 \; \forall \theta \in \Omega_0 \quad (*)$$

y entre todas las pruebas que cumplan $(*)$ se selecciona aquella que maximice la potencia para $\theta \in \Omega_1$.

En nuestro ejemplo suponga que elegimos $\alpha_0 = 0.1$. La región roja indica donde estaría ubicado $\pi(\theta|\delta) \leq \alpha_0$.

```
ggplot() +
    geom_line(
        data = df,
        mapping = aes(x = mu, y = funcion_de_poder, color = tipo), size = 2
    ) +
    geom_rect(
        data = data.frame(xmin = 0, xmax = 8, ymin = 0, ymax = 0.10),
        mapping = aes(xmin = xmin, xmax = xmax, ymin = ymin, ymax = ymax),
        alpha = 0.5, fill = "red"
    ) +
    geom_hline(yintercept = 0.05) +
    theme_minimal()
```



Valor p

Valor p

El Valor p es una herramienta que se puede utilizar para contrastar hipótesis que tiene características que pueden solucionar algunos problemas con los contrastes anteriores. Por ejemplo, no todos los contrastes van a tener un tamaño de contraste exacto (como en los casos donde las variables aleatorias son discretas.) Por otro lado, ni el tamaño ni la potencia están directamente relacionados con los datos observados. El valor p corrige estos problemas. Para propósitos del curso definiremos el valor p como:

> Definición. **Valor p:** Sea $T = T(X_1, X_2, \dots, X_n)$ un estadístico y considere $H_0 : \theta = \theta_0$ contra $H1 : \theta < \theta_0$ (o $H1 : \theta > \theta_0$). Suponga que el contraste rechaza H0 si $T \leq k$ (o $T > k$). Sea $t = T(x_1, x_2, \dots, x_n)$ un valor observado de T. Entonces se define el valor p (denotado pval(t)) como:
>
> $$pval(t) = P(T \leq t | \theta = \theta_0)$$
>
> o
>
> $$pval(t) = P(T > t | \theta = \theta_0)$$

Valor p

Nótese que de esta manera no hay ningún problema si la distribución de T es discreta o continua, pues la probabilidad puede ser calculada sin ningún problema. Formalmente, el valor p se conoce como la probabilidad, bajo la hipótesis nula, de observar nuestro estadístico de prueba o un valor más extremo. Esto significa que si el valor p es grande entonces nuestro estadístico de prueba es un valor muy común de T y se sitúa en el centro de la distribución, brindando así evidencia a favor de la hipótesis nula. Por otra parte, si el valor p es pequeño entonces el estadístico de prueba es un valor poco común de T, situándose en las colas de su distribución, y por lo tanto brindando evidencia en contra de H_0 y a favor de H_1. Existe la posibilidad de que el valor de t haya sido un punto extremo de la distribución de T bajo H_0, pero esto se considera algo como sumamente improbable.

Como altos valores p dan evidencia a favor de H0 y bajos valores dan evidencia en contra, sería posible idear un contraste de hipótesis que consista en rechazar H0 si $pval(t) \leq \alpha$. Si la distribución de T es continua entonces el valor de α sería un tamaño de contraste exacto. En la práctica así es cómo se utiliza el valor p, comparándolo contra el nivel de significancia (tamaño del contraste) y decidiendo si se rechaza o no la hipótesis nula.

Valor p

Ejemplo: Encuentre el valor p para la prueba anterior, suponiendo que $\bar{x} = 5{,}21$ y $n = 4$.

Solución: Recordemos que el contraste consistía en rechazar $H_0 : \mu = 5$ a favor de $H_1 : \mu > 5$ si $x > 5 + \frac{1{,}28}{\sqrt{n}}$. Por lo tanto el valor p consiste en encontrar la siguiente probabilidad:

$$\text{pval}(5{,}21) = P(\bar{X} > 5{,}21 | \mu = 5) = P(Z > \sqrt{4}(5{,}21 - 5)))$$

$$= P(Z > 0,42) = 0,337$$

Por lo tanto si comparamos este valor contra el tamaño de contraste que utilizamos anteriormente (0.10), decimos que no hay suficiente evidencia estadística para rechazar la hipótesis nula de que $\mu = 5$, ya que $\text{pval}(5,21) = 0,337 > 0,10 = \alpha_0$. Esta conclusión es consistente con el contraste que consiste en comparar $\bar{x}$ contra $5 + \frac{1,28}{\sqrt{n}}$. En este caso el valor crítico sería 5.64, por lo que nuestro estadístico de prueba es menor y entonces no rechazamos la hipótesis nula.

Valor p

Existen algunas malas interpretaciones del valor p que no debemos cometer en la práctica. Por lo tanto concluyo esta discusión con dos advertencias:

- El valor p no es la probabilidad de que la hipótesis nula sea cierta. La hipótesis nula es un valor de θ fijo y por lo tanto es cierta o no lo es; no existen probabilidades en este ámbito.
- El valor p no es la probabilidad de cometer un Error Tipo I, dado los datos. El cálculo del valor p no tiene nada que ver con la decisión de aceptar o rechazar la hipótesis nula; es simplemente la probabilidad de una cola de la distribución muestral de T que se puede usar para cuantificar la evidencia de los datos a favor de H0. No obstante, sí se puede usar el valor p para tomar decisiones, pero entonces obviamente en esos caso no sería la probabilidad de Error Tipo I.
- En algunos casos el valor p se conoce como el "tamaño del contraste observado" pues su cálculo es similar al del tamaño del contraste pero utilizando el estadístico de prueba en lugar del valor crítico. No obstante esta es una interpretación que no me gusta pues tiende a confundir la interpretación de un valor p con las interpretaciones erróneas mencionadas anteriormente.

---

class: center, middle

¿Qué discutimos hoy?

Contrastes de hipótesis: función de potencia, tamaño del contraste, el valor p.

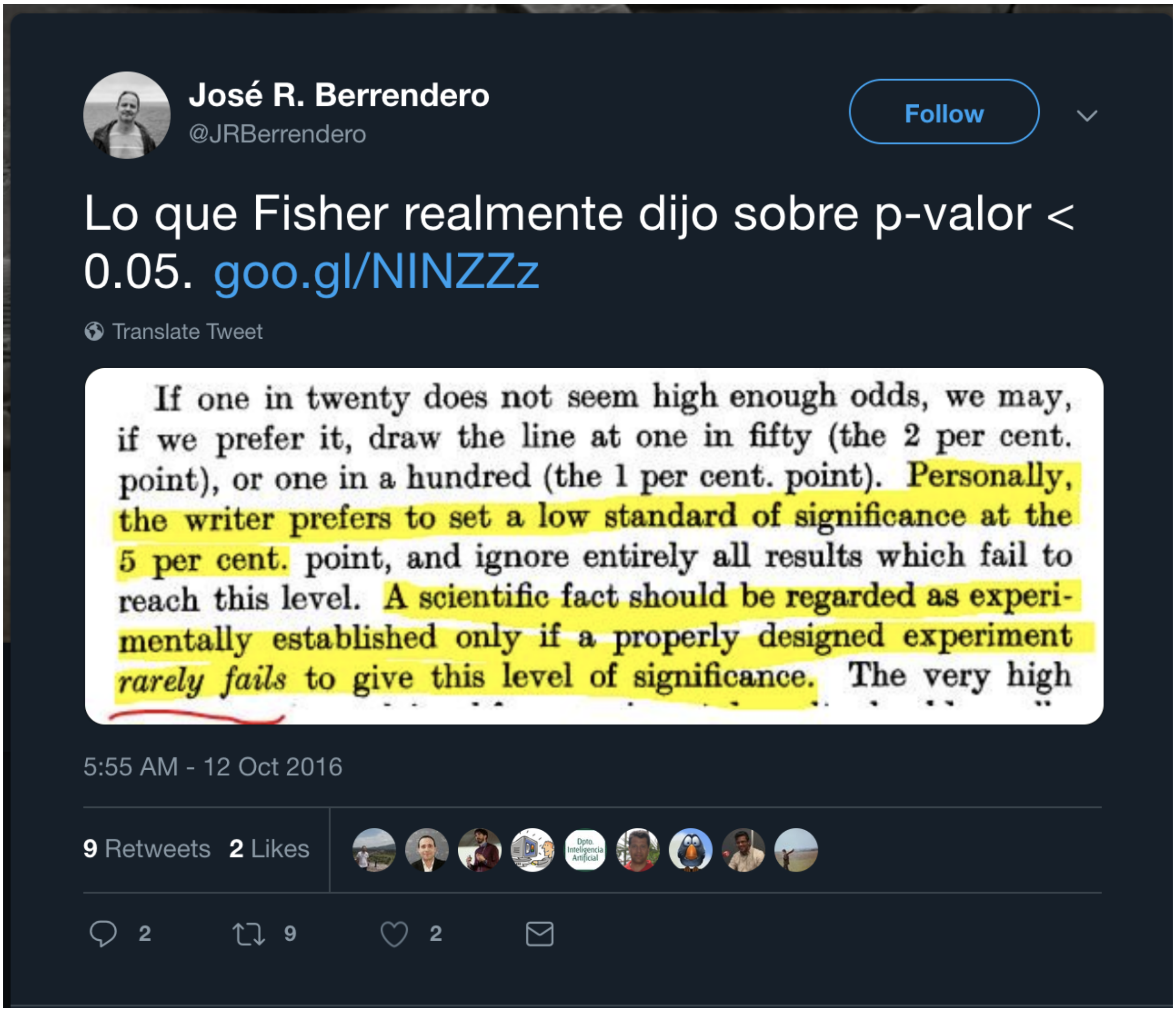


Figure 1: Lo que Fisher dijo de $p < 0.05$