Lec2: High Dimensional Inference

Isidoro Garcia Urquieta

2023

Agenda

- Repaso inferencia estadística

Agenda

- Repaso inferencia estadística
- Inferencia en `Big Data`:

Agenda

- Repaso inferencia estadística
- Inferencia en `Big Data`:
- Ajuste de Bonferroni

Agenda

- Repaso inferencia estadística
- Inferencia en `Big Data`:
- Ajuste de Bonferroni
- False Detectable Rate (FDR)

Repaso de Inferencia Estadística

- La inferencia estadística se refiere a la disciplina de inferir **parámetros** (β) mediante **estadísticos** ($\hat{\beta}_n$) a partir de una muestra de datos de la población de tamaño n.

Repaso de Inferencia Estadística

- La inferencia estadística se refiere a la disciplina de inferir **parámetros** (β) mediante **estadísticos** ($\hat{\beta}_n$) a partir de una muestra de datos de la población de tamaño n.
- Los **parámetros** son números fijos.

Repaso de Inferencia Estadística

- La inferencia estadística se refiere a la disciplina de inferir **parámetros** (β) mediante **estadísticos** ($\hat{\beta}_n$) a partir de una muestra de datos de la población de tamaño n.
- Los **parámetros** son números fijos.
- La **población** es el universo total de unidades a observar i. Típicamente no observamos toda la población.

Repaso de Inferencia Estadística

- La inferencia estadística se refiere a la disciplina de inferir **parámetros** (β) mediante **estadísticos** ($\hat{\beta}_n$) a partir de una muestra de datos de la población de tamaño n.
- Los **parámetros** son números fijos.
- La **población** es el universo total de unidades a observar i. Típicamente no observamos toda la población.
- La **muestra** es una porción de esta población. Esta debe ser **representativa** para poder hacer buenas inferencias.

Repaso de Inferencia Estadística

- La inferencia estadística se refiere a la disciplina de inferir **parámetros** (β) mediante **estadísticos** ($\hat{\beta}_n$) a partir de una muestra de datos de la población de tamaño n.
- Los **parámetros** son números fijos.
- La **población** es el universo total de unidades a observar i. Típicamente no observamos toda la población.
- La **muestra** es una porción de esta población. Esta debe ser **representativa** para poder hacer buenas inferencias.
- Los **estadísticos** se construyen a partir de los datos observados. Ellos estiman un aproximado del parámetro. Es por eso que son **variables aleatorias**.

Repaso de Inferencia Estadística

- La inferencia estadística se refiere a la disciplina de inferir **parámetros** (β) mediante **estadísticos** ($\hat{\beta}_n$) a partir de una muestra de datos de la población de tamaño n.
- Los **parámetros** son números fijos.
- La **población** es el universo total de unidades a observar i. Típicamente no observamos toda la población.
- La **muestra** es una porción de esta población. Esta debe ser **representativa** para poder hacer buenas inferencias.
- Los **estadísticos** se construyen a partir de los datos observados. Ellos estiman un aproximado del parámetro. Es por eso que son **variables aleatorias**.
- Asi, un estimador $\hat{\beta}_n$ va a tener distintos valores para muestras distintas de la población $(n_1, n_2, n_3, ...)$

Algunas propiedades de los estadísticos:

(In)Sesgo: Si promediaramos los estimadores de todas las muestras, que tan cerca o lejos estaríamos del parámetro poblacional

$$Sesgo(\hat{\beta}_n) = E(\hat{\beta}_n) - \beta$$

Error: Cual es la diferencia entre un estimador en particular vs el parámetro?

$$e = \hat{\beta}_n - \beta$$

Varianza: Que tanto difiere cada estimador del promedio de los estimadores (al cuadrado)

$$var(\hat{\beta}_n) = E(\hat{\beta}_n - E[\hat{\beta}_n])^2$$

Error Cuadrâtico Medio: Esto es el promedio del los errores al cuadrado.

$$ECM(\hat{\beta}_n) = E[(\hat{\beta}_n - \beta)^2] = var(\hat{\beta}_n) + Sesgo(\hat{\beta}_n)^2$$

En general quieres estimadores que sean insesgados y/o que tengan mínima varianza (i.e. El menor *ECM*). En la econometría (siguiente sesión) se enfatiza la **insesgadez**. Veremos que hay un bias+variance tradeoff.

Otras propiedades del comportamiento de los estimadores:

Eficiencia: Queremos el estimador con el menor *ECM*. Por ejemplo, en el mundo de estimadores insesgados, buscamos el que tenga menor varianza.

Consistencia: Si el tamaño de la muestra n crece mucho ($n \to \infty$), el estimador converge al párametro.

$$lim_{n\to\infty} Pr[|\hat{\beta}_n - \beta| < \epsilon] = 1$$

o

$$\hat{\beta}_n \to^p \beta$$

Normalidad: La distribución de los estimadores es normal alrededor del parámetro.

$$\frac{\hat{\beta}_n - \beta}{\sqrt{\frac{\sigma^2}{n}}} \to^D N(0,1)$$

Ejemplo: Supongamos que tienes una población de 100,000 individuos cuya edad promedio es ~23 años. Quieres construir un estimador a partir de muestras de 15,000 personas.

El parámetro es $\mu_{edad} = \frac{1}{100,000}\Sigma_{i=1}^{100,000} edad_i = 23$

El estimador más obvio sería: $\hat{\mu_{edad}} = \frac{1}{15,000}\Sigma_{i=1}^{15,000} edad_i = X$

El valor del estimador depende de la **muestra aleatoria** que tomemos! Veamos como se ve en R

```
library(tidyverse)
set.seed(1957)
poblacion<-tibble(id = seq(1:100000),
                  edad = rchisq(n = 100000,df = 5)+18)

(parametro<-mean(poblacion$edad))
```

```
## [1] 22.98495
```

```r
# Tomamos 100 muestras de 15000 observaciones
muestras<-map(1:100,
              function(x) {
                  set.seed(x)
                  poblacion %>%
                      slice_sample(n = 15000, replace = T)
              })

head(map(muestras, ~dim(.)))
```

```
## [[1]]
## [1] 15000     2
##
## [[2]]
## [1] 15000     2
##
## [[3]]
## [1] 15000     2
##
## [[4]]
## [1] 15000     2
```

Ahora construyamos el estimador $\hat{\mu_{15,000}}$ para cada muestra:

```
estimadores<-map_dbl(muestras, ~ mean(.$edad))

estimadores<-tibble(muestra = seq(1:100),
                    estimador = estimadores)
```

Veamos como se ven los 100 estimadores vs el parametro 22.98

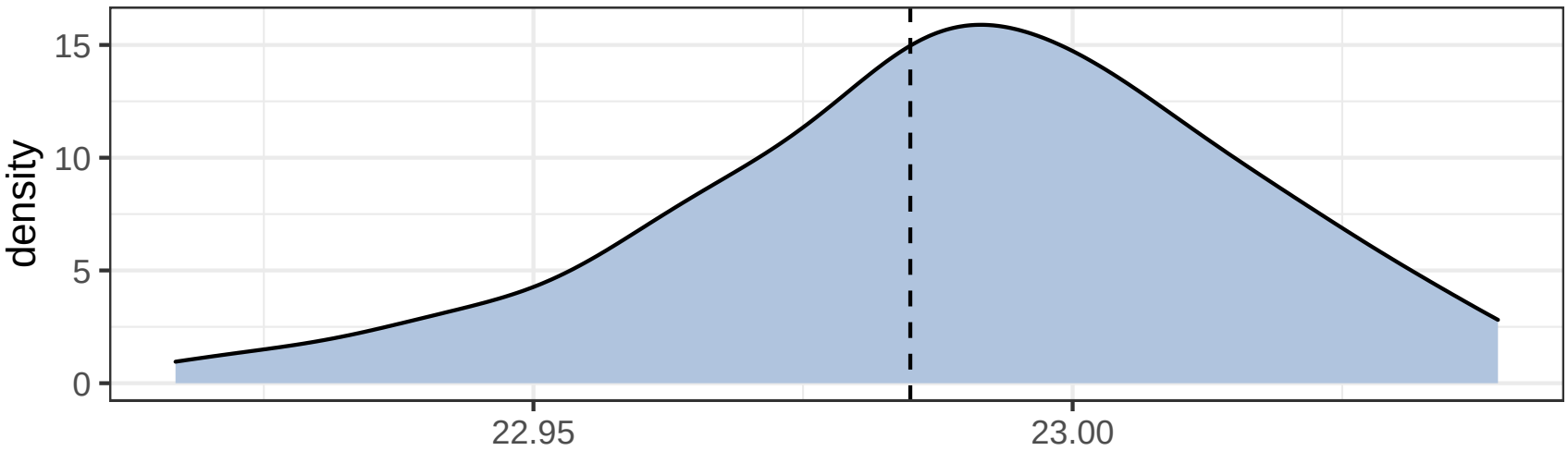
```
(media<-mean(estimadores$estimador))
```

```
## [1] 22.98937
```

```
(varianza<-var(estimadores$estimador))
```

```
## [1] 0.0006647197
```

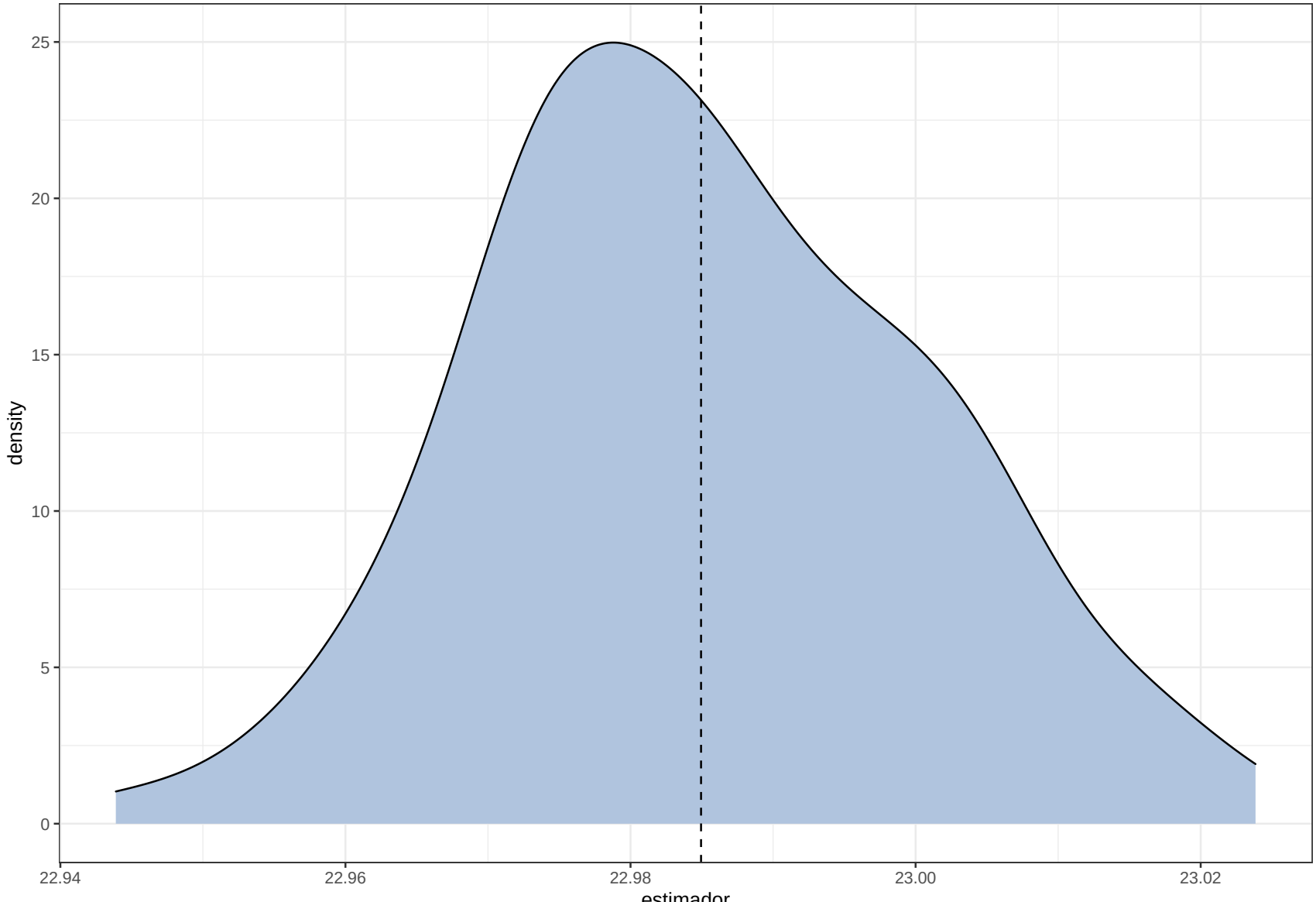
```
ggplot(estimadores, aes(estimador))+
   geom_density(fill = 'lightsteelblue')+theme_bw()+
    geom_vline(xintercept = mean(poblacion$edad),
               linetype = 'dashed')
```



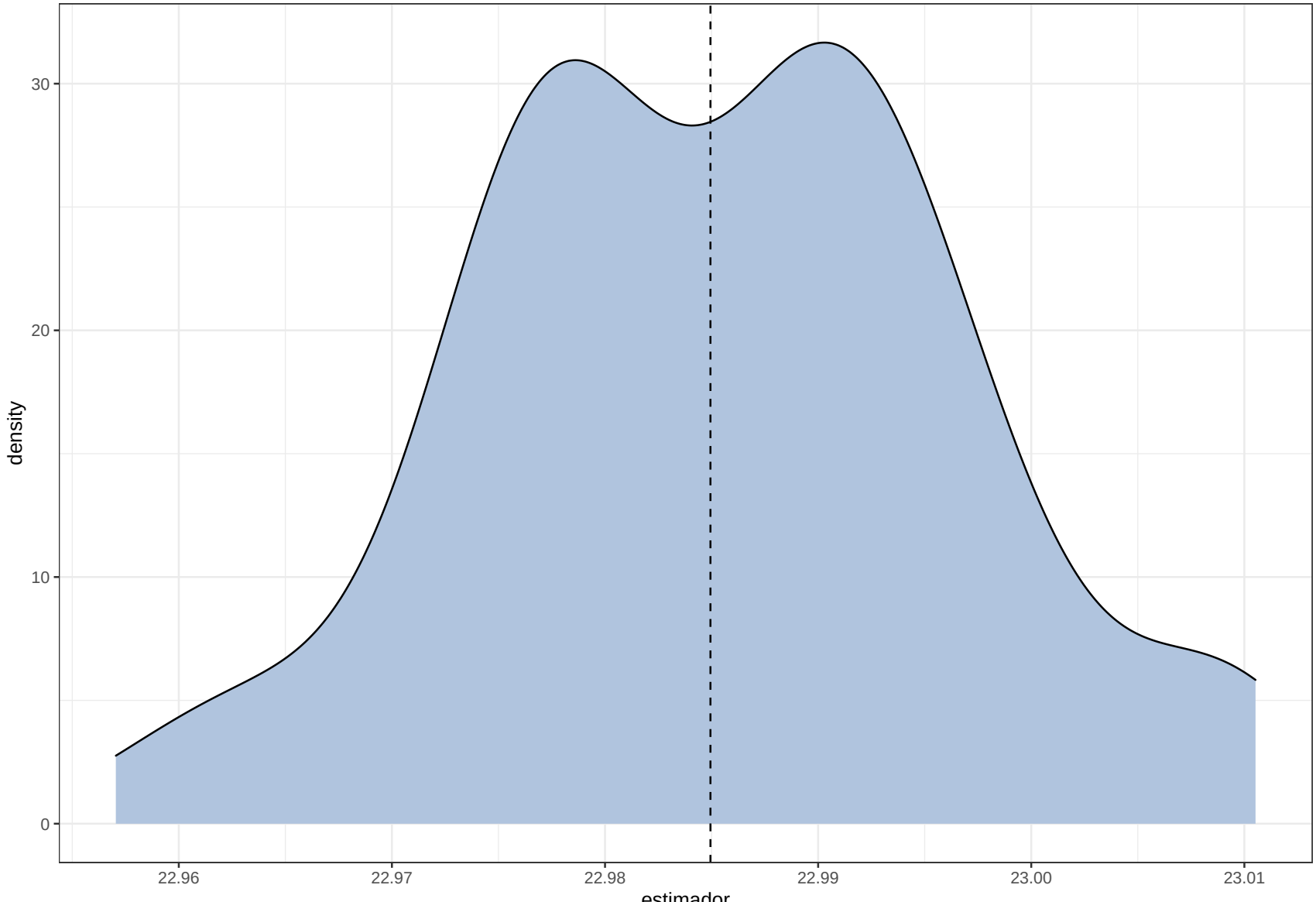
Notemos algunas cosas:

- $Sesgo(\hat{\beta}) = 22.9803 - 22.9827 = -0.0024$
- $var(\hat{\beta}) = 0.0006785$
- Hay 100 **errores**, uno por cada estimador
- $ECM(\hat{\beta}) = var(\hat{\beta}_n) + Sesgo(\hat{\beta}_n)^2 = 0.0006842804$
- Es un estimador muy bueno! Poquísimo sesgo, baja varianza.
- Se cumple la **normalidad**: Noten como $edad \sim \chi_5$ y el estimador igual es normal.
- Nos falta mostrar la consistencia! Recordemos, si hacemos la muestra mas grande, la distribucion debe colapsarse al parametro.

Para 45,000 observaciones



Para 90,000 observaciones



Nota en Normalidad

Vean como mostramos que el **estimador**:

$$\hat{\beta}_n \rightarrow^D N(\beta, \sqrt{\sigma^2/n})$$

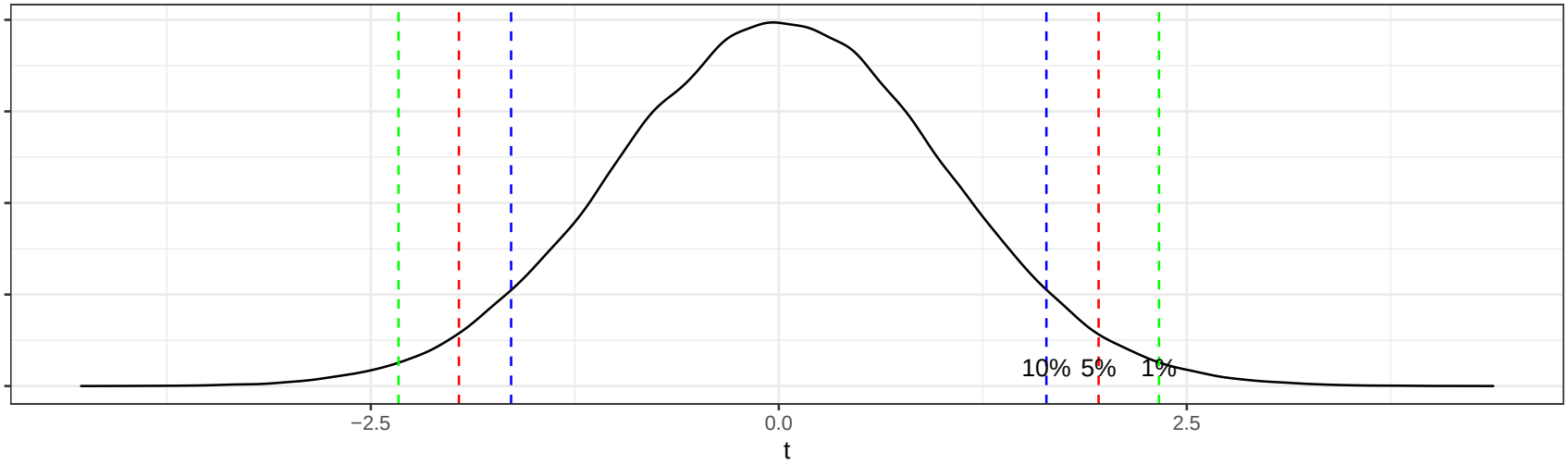
Y esto es lo mismo que (**estimador estandarizado**):

$$\frac{\hat{\beta}_n - \beta}{\sqrt{\frac{\sigma^2}{n}}} \rightarrow^D N(0,1)$$

Pruebas de Hipótesis

Las pruebas de hipótesis ven el estimador estandarizado $t = \frac{\hat{\beta}_n - \beta_{H0}}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$ de una muestra en particular para inferir (intentar probar) el valor de un parámetro.

Esto es, dado el estimador, que tan probable es que nuestra H_0 sea correcta?



Pruebas de Hipótesis

Los conceptos más importantes son:

- Hipótesis Nula (H_0): Valor que quieres rechazar/no rechazar
- Hipótesis Alternativa (H_a): Un rango del valor. Una cola o dos colas
- Nivel de significancia ($\alpha = \Pr[|t|>Z^*] = Pr[Rech \ \ H_0|H_0 \ V]$)
- Estadístico ó Estimador (t)
- Valor crítico (Z^*): Punto(s) de corte para la significancia elegida
- P-value: Probabilidad de observar el valor del t bajo H_0
- Poder: $Pr[Rech \ \ H_0|H_0 \ F]$
- **Problema en Big data**: cuanto mas grande la n, mas grande $t = \frac{\hat{\beta}_n - \beta_{H0}}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$. Esto hace que demos muchas pruebas por significativas.

Ejemplo

Volvamos a nuestro ejemplo de la población de 100,000 con una edad promedio de ~23.

Tomamos **una** muestra de 45,000.

```
library(tidyverse)
set.seed(1957)
poblacion<-tibble(id = seq(1:100000),
                  edad = rchisq(n = 100000,df = 5)+18)

(parametro<-mean(poblacion$edad))
```

```
## [1] 22.98495
```

```r
# Tomamos 1 muestra de 45000 observaciones
set.seed(1990)
muestra<-
   poblacion %>%
   slice_sample(n = 45000, replace = T)

# Estimadores

# Media
(m<-mean(muestra$edad))
```

```
## [1] 22.99009
```

```r
# Varianza
(v<-var(muestra$edad))
```

```
## [1] 9.946144
```

Ejemplo

Ahora hagamos una hipótesis simple: La edad promedio de la población es 30 años? (i.e. $\beta_{H_0} = 30$)

$$H_0 : \beta = 30 \quad H_a : \beta \neq 30$$

El estadístico de prueba es: $\frac{\hat{\beta}_n - \beta_{H_0}}{\sqrt{\frac{\sigma^2}{n}}} = \frac{22.99009 - 30}{\sqrt{\frac{9.946144}{45,000}}} = -471.5104$

```
# El p-value
(est<-sqrt(45000)*(m -30)/sqrt(v))
```

```
## [1] -471.5104
```

```
p_value<-pnorm(q = est)
```

$P[Rechazar \;\; H_0 | \beta = 30 \;\; (H_0 \;\; verdadera)] = 0$. Este $p - value$ es mucho menor que los cortes de $(0.01, 0.05, 0.1)$. Es decir, rechazamos que la edad promedio sea 30 años y tenemos una probablilidad de 0 de equivocarnos.

Pruebas conjuntas

En las pruebas conjuntas incluimos varios parametros a la vez. Es muy común para evaluar si una regresión explica mas que la media simple de y (i.e. le gana al modelo nulo).

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \qquad H_a : \beta_i \neq 0$$

El estadístico que se usa es el $F = \frac{SSE/k}{SSR/(n-k-1)}$

Noten como en esta prueba también el estadístico de prueba F es **creciente** en **n**. Es decir, en un mundo de **Big Data** sería demasiado fácil incluir muchas variables en un modelo; aún si estas no fueran realmente relevantes.

Muchas pruebas en Big Data

Supongamos que tenemos una base de dimensiones $[n, p]$. Queremos hacer pruebas de significancia para cada variable p:

$$H_{01}, H_{02}, H_{03}, ..., H_{0p} : \beta_{i \in p} = 0$$

Supongamos ademas que en $N_0 < p$ de esas pruebas la hipótesis nula es verdadera. Para $N_1 = p - N_0$ la hipótesis nula es falsa.

Veamos los errores:

Real	No_Rechazar	Rechazar
Noise	Negativos Verdaderos	Falsos Positivos
Signal	Falsos Negativos	Positivos verdaderos

Queremos un algoritmo que balanceé ambos errores.

El problema de muchas columnas

Para una significancia α, si corremos muchas pruebas, α por ciento de ellas saldrán significativas sólo por azar.

Veamos un ejemplo de regresión con 100 variables, 5 de ellas realmente significativas. Mas aún, las 5 de ellas las encuentras significativas.

Para las 95 variables restantes (todas ruido), corres pruebas a $\alpha = 0.05$. Encontrarías 4.75 variables significantes cuando no lo son (falsos positivos). Más aún, $\frac{4.75}{4.75+5} \approx 50$ de nuestros positivos son falsos!!

Esto es el False Discovery Proportion (FDP)

Noten como depende de la cantidad de verdaderos positivos y de α.

False Discovery Rate

Noten como no podemos conocer el *FDP*, pues depende de los desconocidos verdaderos positivo. *FDP* es un parámetro!

$$FDP = \frac{Total \quad falsos \quad positivos}{tests \quad significantes}$$

Afortunadamente, existe un estimador para *FDP*, False Discovery Rate, $FDR = E[FDP]$.

Este es un análogo multivariado de α te ayuda a controlar que tu modelo no tenga demasiados falsos positivos.

False Discovery Rate

Con la inferencia tradicional, elegimos α y a partir de ahí se genera un *FDR*: $q(\alpha)$

Con FDR control, fijamos un nivel de $FDR \leq q$ y de ahí se genera un nuevo corte de significancia: $\alpha(q)$

Algortimo Benjamini + Hochberg:

- Eliges un nivel de FDR q (por ejemplo 0.1)
- Rankeas tus N p-values de menor a mayor. El ranking de cada p-value se guarda en un vector k (i.e. el p-value más pequeño es $k = 1$)
- Elige el nuevo corte $p* = max\{p(k) : p(k) \leq q\frac{k}{N}\}$
- Rechazas las H_0 con $p \leq p* \rightarrow$ tu $FDR \leq q$

Importante: FDR control asume que las pruebas son indenpendientes entre ellas.

False Discovery Rate

Veamos de vuelta la regla:

$$p* = max\{p(k) : p(k) \leq q\frac{k}{N}\}$$

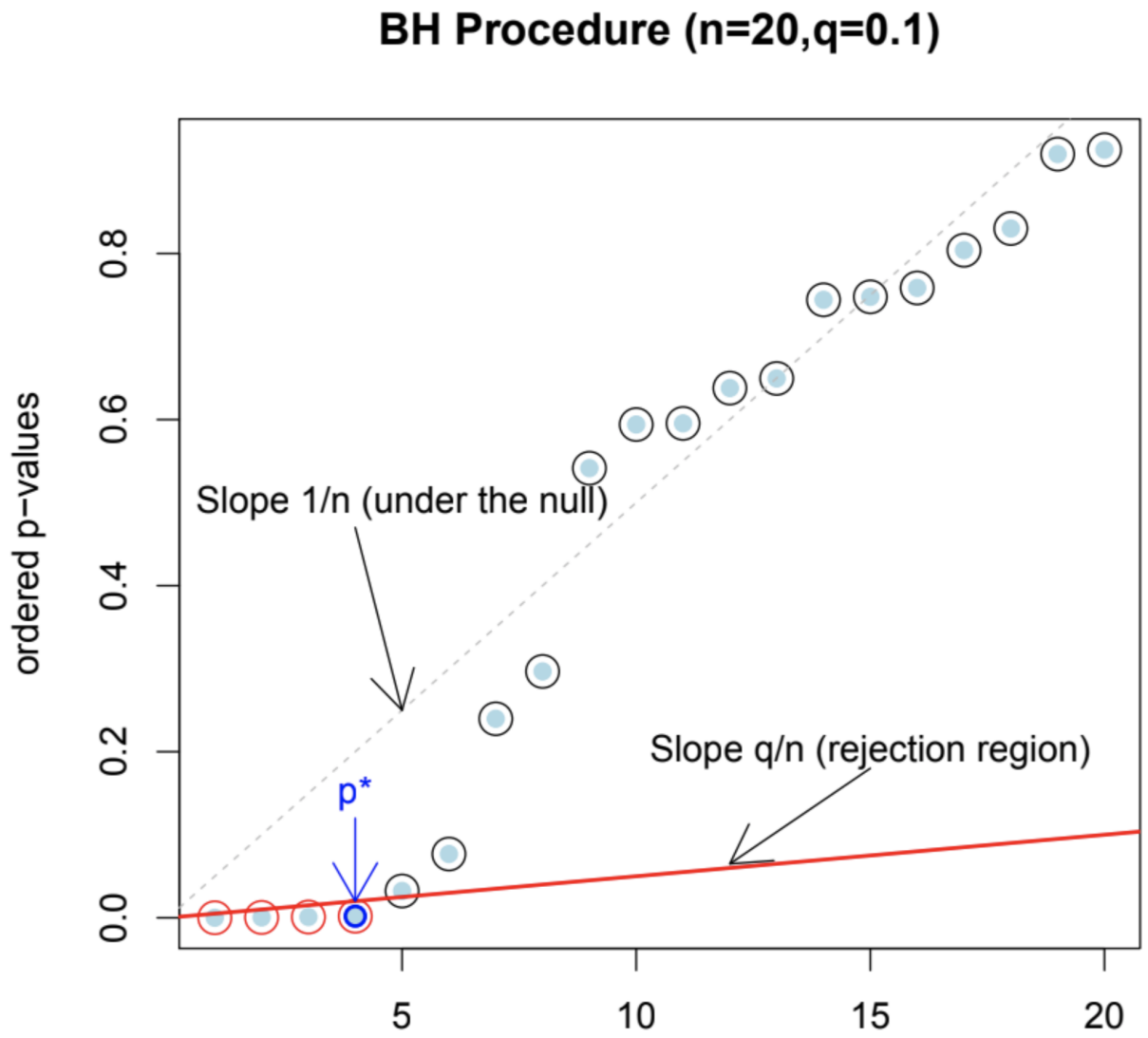
Esto es, eliges el máximo de tus $p - value$ cuyo valor es menor al valor proporcional por su ranking.

Ejemplo: Imagina corres una regresión con $N = 100$ variables. Quieres que sólo 10% de tus variables significativas sean falsos positivos (i.e. $q = 0.1$).

Con esto, el $p - value$ más pequeño ($p(1)$) debe ser menor a $p(1) \leq 0.1\frac{1}{100} = 0.001$ para 'pasar'. Tu p-value con ranking 80 $p(80)$ debe ser más pequeño que $p(80) \leq 0.1\frac{80}{100} = 0.08$ para entrar. Así sucesivamente.

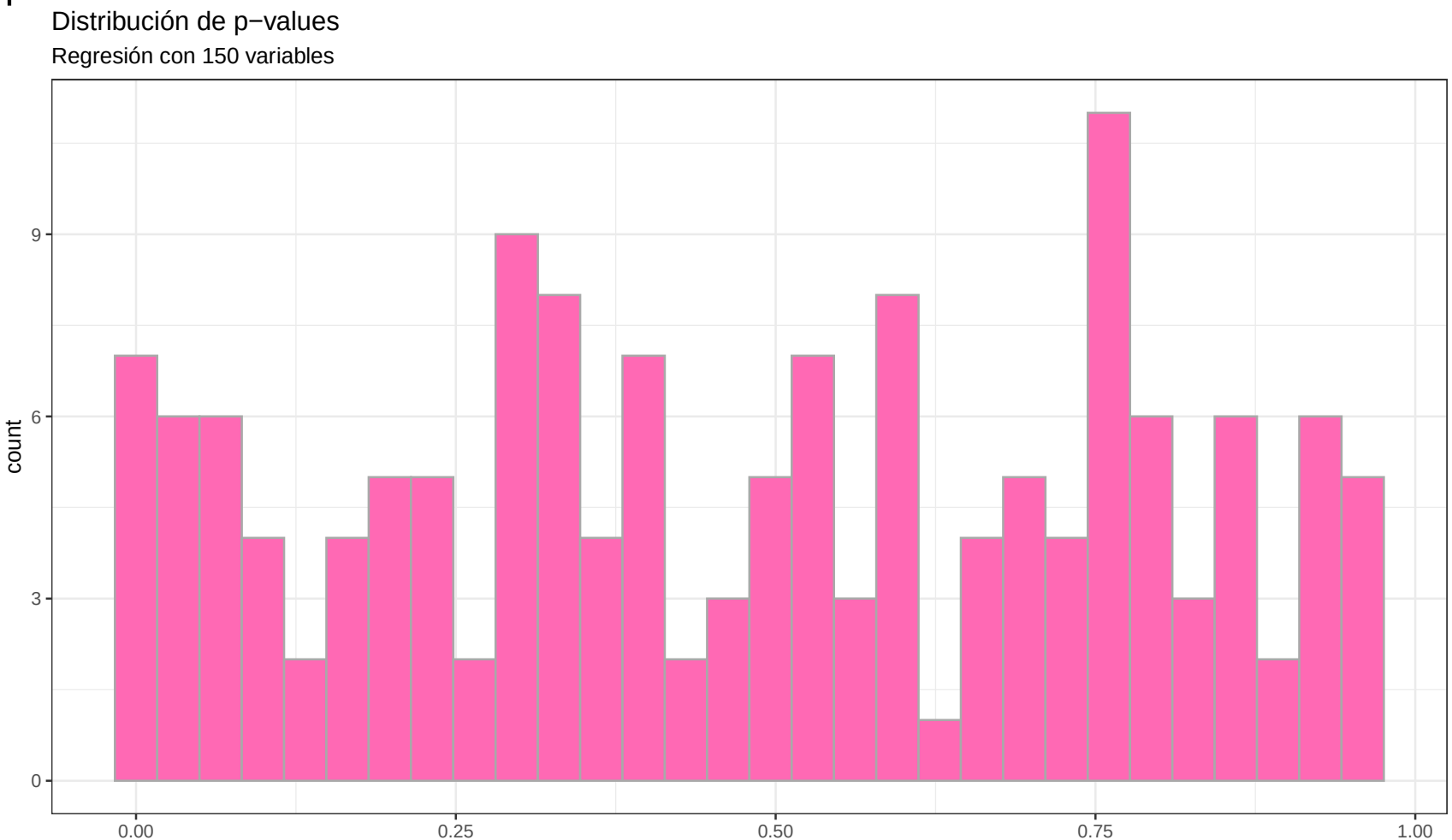
El algoritmo de Benjamini simplemente escoge el mayor (último p-value) que cumple con está regla y lo elige como corte de significancia α.

False Discovery Rate

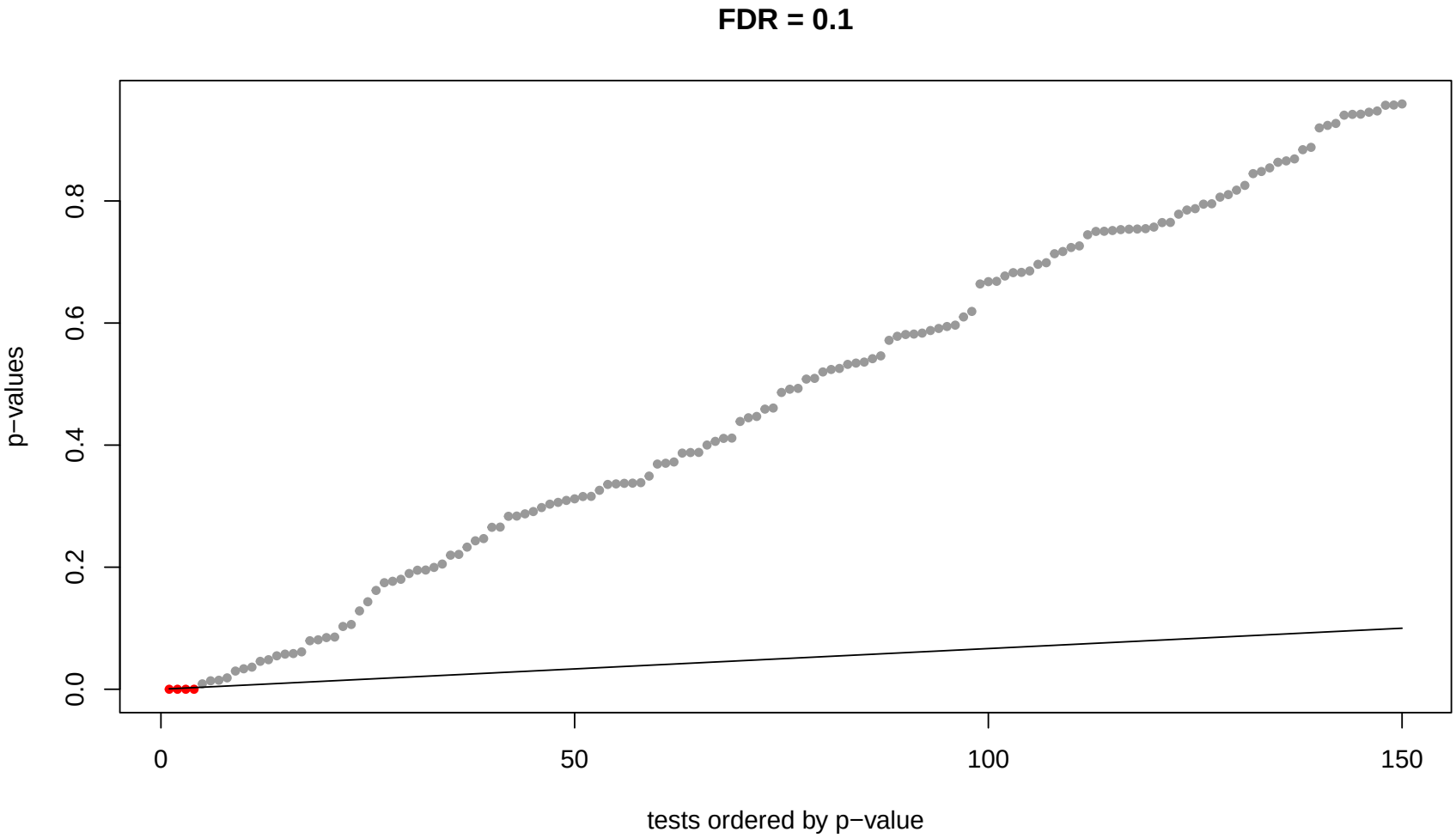


Ejemplo

Esta es la distribucion de $p - values$ de una regresión de 150 variables. Cuántas son significativas si queremos una tasa de falsos positivos de 10 por ciento?



Ejemplo




```
## [1] 1e-05
```