

Fundamentos de Inferencia Estadística

Francisco Javier Mercader Martínez

Índice

1	Muestreo y distribuciones muestrales	2
1.1	Introducción	2
1.2	Ejemplos	2
1.3	Surge una pregunta	2
1.4	Esbozo de respuesta: tasa de participación	3
1.5	Realización del experimento: conclusiones	4
1.6	En la práctica	4
1.7	Uso de la distribución muestral	5
1.8	Antes de extraer una muestra:	5
1.9	Otro ejemplo: valores muestrales de una distribución normal	5
1.10	Un resultado importante	6
1.11	Algunos términos	6
1.12	Ejemplos de estadísticos	6
1.13	La media muestral	7
1.13.1	Esperanza y varianza de la media muestral	7
1.14	Consecuencia práctica	8
1.14.1	Analogía con una diana	8
1.15	Varianza muestral	8
1.15.1	Dos apuntes	9
1.16	Esperanza de la varianza muestral	9
1.17	Distribuciones muestrales de \bar{X} y S^2	9
1.18	Distribución de \bar{X} y S^2 para una m.a.s. de una distribución normal	9
1.19	Recordatorio: distribución χ^2 con p grados de libertad	9
1.20	Distribución t-Student	10
1.20.1	¿Cuál es la forma de la densidad de una t-Student?	11

Tema 1: Muestreo y distribuciones muestrales

1.1) Introducción

i El contexto

- Tenemos una pregunta acerca de un fenómeno aleatorio.
- Formulamos un modelo para la variable de interés X .
- Traducimos la pregunta de interés en términos de uno o varios parámetros del modelo.
- Repetimos el experimento varias veces, apuntamos los valores de X .
- ¿Cómo usar estos valores para extraer información sobre el parámetro?

1.2) Ejemplos

i ¿Está la moneda trucada?

- Experimento: tirar una moneda. X = resultado obtenido:

$$P(X = +) = p, P(X = c) = 1 - p$$

$$¿p = \frac{1}{2}?$$

i Sondeo sobre intención de participación en una elecciones

- Queremos estimar la tasa de participación antes de unas elecciones generales.
- Formulamos un modelo:
 - Experimento: "escoger una persona al azar en el censo".
 - X : participación, variable dicotómica ("Sí" o "No"). $p = P(X = \text{"Sí"})$.
- ¿Cuánto vale p ?
- Censo: aprox. 37 000 000. Escogemos aprox. 3000 personas.

i Determinación de la concentración de un producto

- Quiero determinar la concentración de un producto.
- Formulo el modelo:
 - Experimento: "llevar a cabo una medicación".
 - X : "valor proporcionado por el aparato".
 - $X \sim \mathcal{N}(\mu, \sigma^2)$.
- ¿Qué vale μ ?

1.3) Surge una pregunta

En todas estas situaciones donde nos basamos en la repetición de un experimento simple...

- ¿Cómo sabemos que nuestra situación es fiable?
- ¿Qué confianza tenemos al extrapolar los resultados de una muestra de 3000 personas a una población de 37 millones de personas?

1.4) Esbozo de respuesta: tasa de participación

i Para convenceros, un experimento de simulación

- Voy a simular el proceso de extracción de una muestra de 3000 personas en una población de 37 millones de personas.
- Construyo a mi antojo los distintos componentes:
 - **La población:** defino en mi ordenador un conjunto de 37 000 000 de ceros y unos (\Leftrightarrow el censo electoral)
 - "1" \Leftrightarrow "la persona piensa ir a votar".
 - "0" \Leftrightarrow "la persona no piensa ir a votar".
 - **La tasa de participación "real":** Decido que en mi población el 70% piensa en ir a votar \rightarrow 25 900 000 "1"s.
 - **La extracción de una muestra:** construyo un pequeño programa que extrae al azar una muestra de 3000 números dentro del conjunto grande.

```
## [1] 0.7056667
```

Queremos descartar que haya sido suerte. Vamos a repetir muchas veces (10000 veces por ejemplo), la extracción de una muestra de 3000 personas en la población.

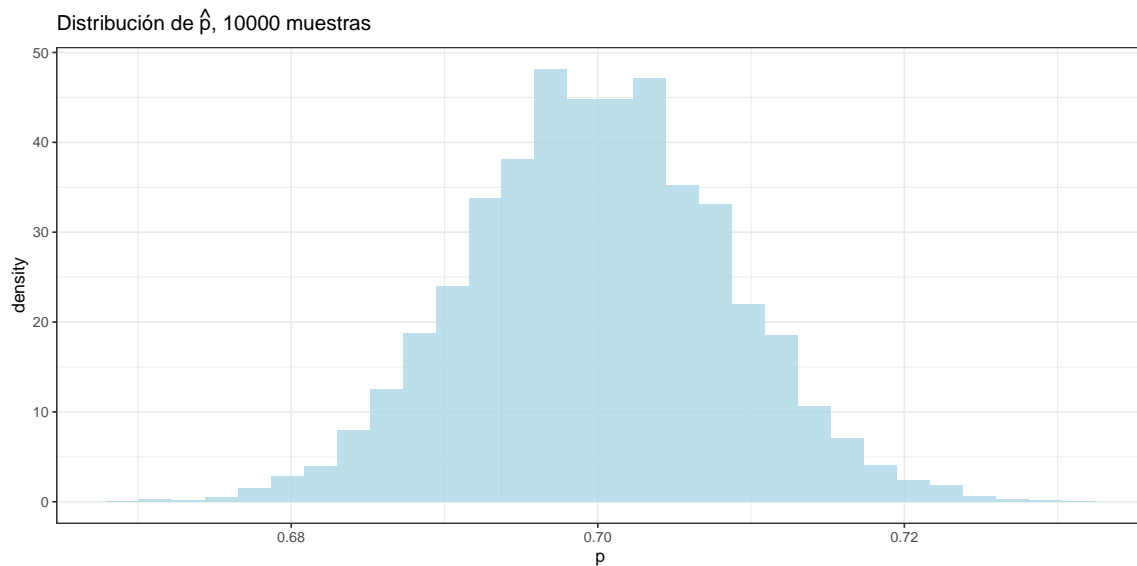
```
library(tidyverse)
lista_muestras <- replicate(
  10000,
  sample(poblacion, 3000, replace = FALSE),
  simplify = FALSE
)
p_muestras <- map_dbl(lista_muestras, mean)
head(p_muestras)

## [1] 0.6970000 0.7030000 0.7036667 0.7023333 0.7013333 0.7226667
```

```
library(tidyverse)
p_muestras <- replicate(
  10000,
  sample(poblacion, 3000, replace = FALSE),
  simplify = FALSE
) |>
  map_dbl(mean)
head(p_muestras)

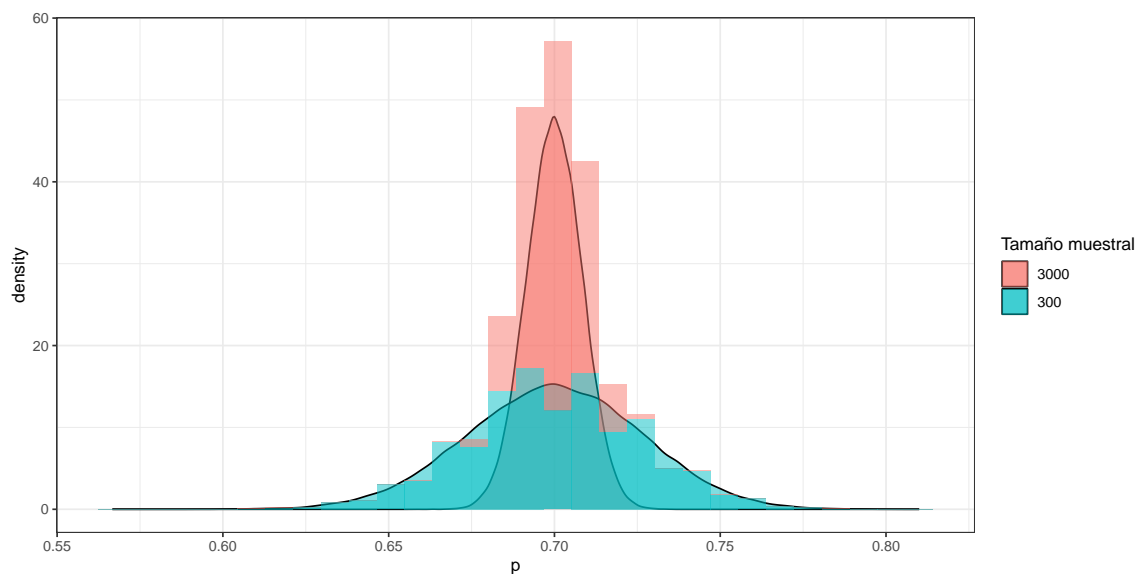
## [1] 0.7090000 0.7080000 0.6946667 0.7216667 0.6906667 0.6990000
```

Recogemos los valores obtenidos en un histograma.



1.5) Realización del experimento: conclusiones

- La enorme mayoría de las muestras de 300 individuos proporcionan una tasa de participación muy próxima a la de la población.
→ **El riesgo** de cometer un error superior a ± 2 puntos, al coger una **una** muestra de 3000 individuos es muy pequeño (y asumible...)
- Si nos limitamos a muestras de 300 individuos, ¿qué esperarías?



1.6) En la práctica

i Usamos las distribuciones muestrales

- Las empresas de sondeos no se basan en simulaciones sino en cálculos teóricos.
- Experimento aleatorio: escoger al azar una muestra de 3000 personas dentro de una población de 37 000 000, con una tasa de participación p .
- Llamamos a \hat{p} la variable aleatoria: proporción de "1"s en la muestra escogida.
- ¿Cuál es la distribución de valores de \hat{p} ?

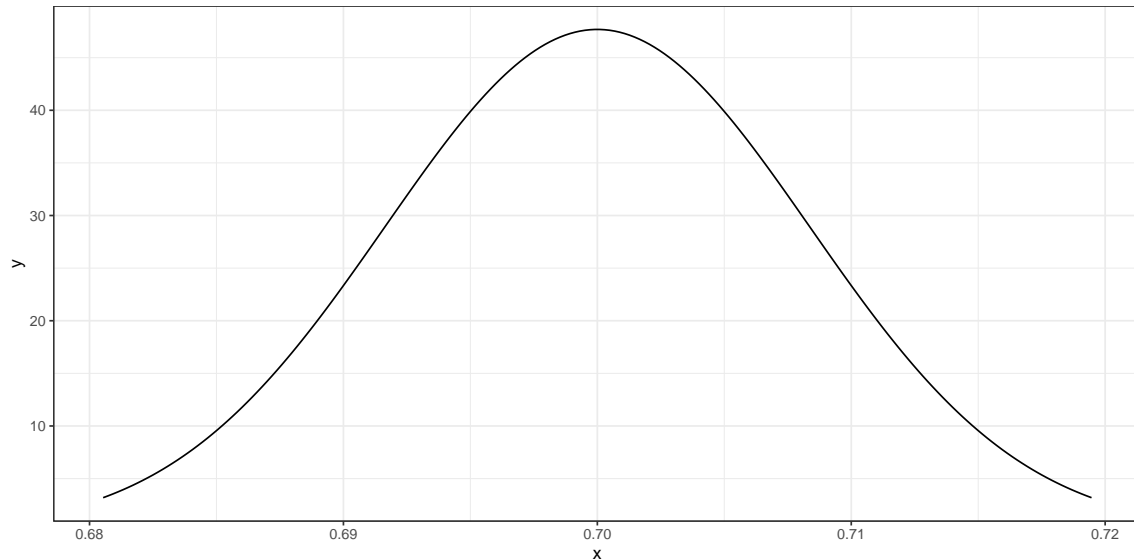
$$\hat{p} \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$$

Es lo que llamamos la **distribución muestral** de \hat{p} .

1.7) Uso de la distribución muestral

i La distribución muestral de \hat{p} :

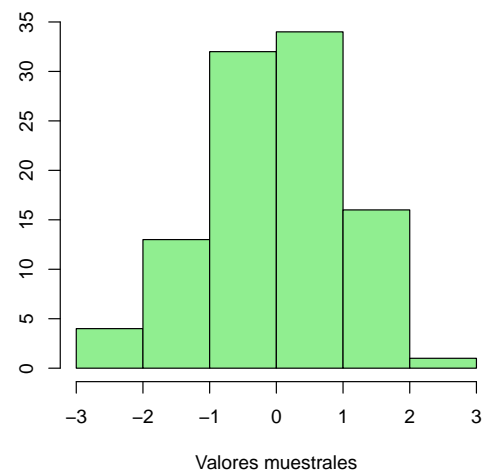
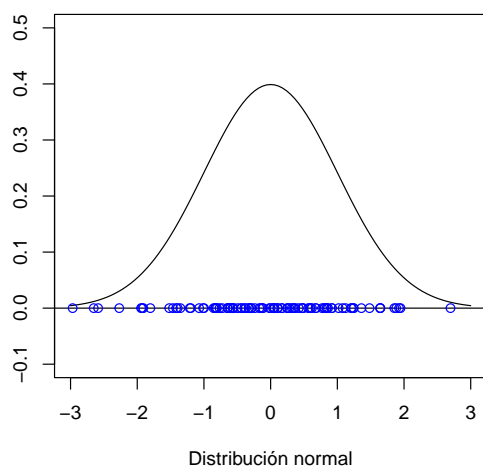
Es la distribución esperada de los valores de \hat{p} respecto a todas las muestras de ese tamaño que podría extraer

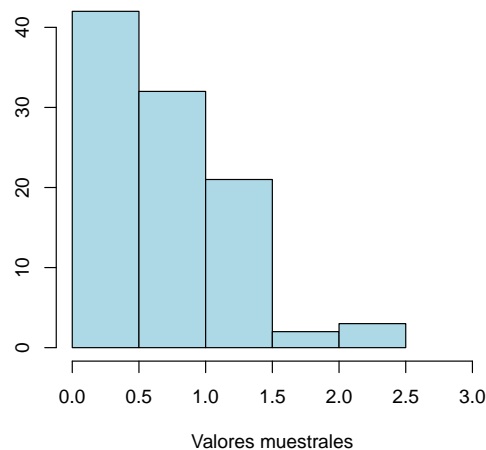
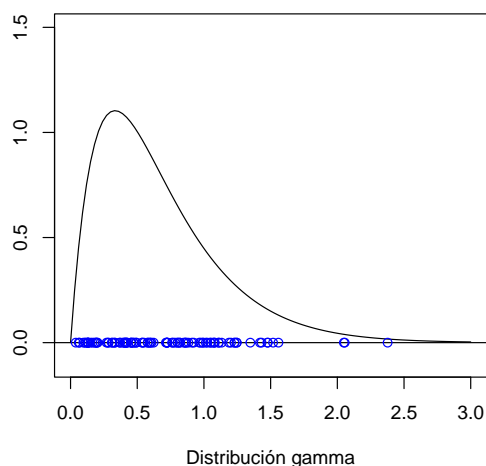


1.8) Antes de extraer una muestra:

- ¿Es suficiente el tamaño para el riesgo asumible y la precisión requerida?
- Una vez la muestra:
 - ¿Puedo dar un margen de error?
 - ¿Puedo decidir si p poblacional es, por ejemplo, mayor que un valor dado?

1.9) Otro ejemplo: valores muestrales de una distribución normal





1.10) Un resultado importante

i Ley (débil) de los grandes números

Sea X una variable aleatoria y $g(X)$ una variable aleatoria transformada de X , con esperanza y momento de orden 2 finitos. Supongamos $X_1, X_2, \dots, X_n, \dots$ una sucesión de variables aleatorias (v.v.aa) independientes con la misma distribución que X , entonces

$$\lim_{n \rightarrow \infty} P \left[\left| \frac{\sum_{i=1}^n g(X_i)}{n} - E[g(X)] \right| < \varepsilon \right] = 1, \text{ para todo } \varepsilon > 0.$$

1.11) Algunos términos

Definición 1.11.1 • Sea una variable aleatoria X . Consideramos n variables aleatorias independientes e idénticamente distribuidas X_1, X_2, \dots, X_n , que se distribuyen como X . La variable aleatoria multidimensional (X_1, X_2, \dots, X_n) es una *muestra aleatoria simple* (m.a.s.) de X .

- Cualquier cantidad calculada a partir de las observaciones de una muestra: *estadístico*.
- Experimento aleatorio: extraer una muestra. Consideramos un estadístico como una variable aleatoria. Nos interesa conocer la distribución del estadístico: *distribución muestral*.

1.12) Ejemplos de estadísticos

- Proporción muestral: \hat{p} .
- Media muestral: $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$.
- Desviación típica muestral: $S_X = \sqrt{\frac{1}{n+1} \sum_{i=1}^n (X_i - \bar{X})^2}$

1.13) La media muestral

i Contexto

Estudiamos una variable X cuantitativa.

- Estamos interesados en μ , el centro de la distribución de X .
- Extraemos una muestra de tamaño n :

$$x_1, x_2, \dots, x_n.$$

- Calculamos su media \bar{x} para aproximar μ .
- ¿Cuál es la distribución muestral de \bar{X} ?

Ejemplo:

- Quiero medir una cantidad. Hay variabilidad en las mediciones.
- Introduzco una variable aleatoria X = "valor proporcionado por el aparato".
- μ representa el centro de los valores.
- Extraigo una muestra de tamaño 5 del valor de X .

1.13.1) Esperanza y varianza de la media muestral

Llamamos $\mu = E[X]$ y $\sigma^2 = \text{Var}(X)$.

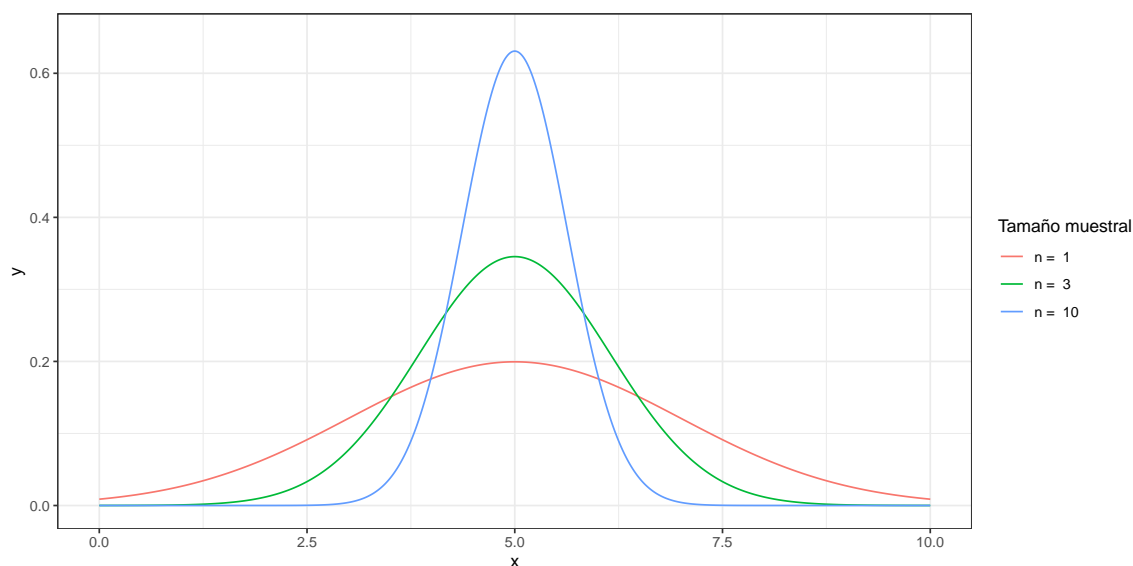
- Tenemos

$$E[\bar{X}] = \mu.$$

→ Es decir que el centro de la distribución muestral de \bar{X} coincide con el centro de la distribución X .

- Tenemos $\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$, es decir, la dispersión de la distribución muestral de \bar{X} es \sqrt{n} veces más pequeña que la dispersión inicial de X .

Ilustración: X inicial, \bar{X} con $n = 3$, \bar{X} con $n = 10$.

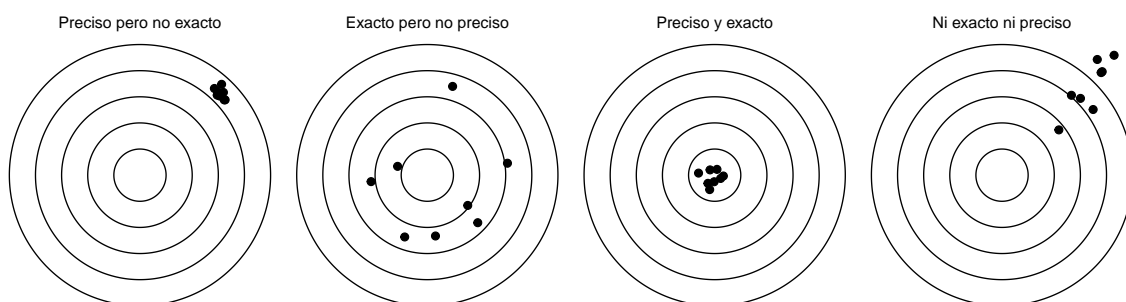


1.14) Consecuencia práctica

i Aparato de medición

- Experimento: llevar a cabo una medición con un aparato.
- Variable aleatoria X : "valor proporcionado por el aparato".
- $E[X]$: centro de la distribución de los valores proporcionados por el aparato.
 - Lo deseable: $E[X]$ = valor exacto de la cantidad que buscamos medir.
 - En este caso, decimos: el aparato es **exacto**.
- σ_X : dispersión de la distribución de los valores proporcionados por el aparato.
 - Lo deseable: σ_X pequeño.
 - En este caso, decimos: el aparato es **preciso**.

1.14.1) Analogía con una diana



1.15) Varianza muestral

Si (X_1, X_2, \dots, X_n) es una muestra aleatoria simple de X , definimos la **varianza muestral** S_n^2 como

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

i Fórmula alternativa para S_n^2 :

$$S_n^2 = \frac{n}{n-1} \left(\overline{X^2}_n - (\bar{X}_n)^2 \right),$$

donde $\overline{X^2}_n = \frac{1}{n} \sum_{i=1}^n X_i^2$.

1.15.1) Dos apuntes

i En algunos textos en castellano

Se suele llamar S_n^2 **cuasi-varianza muestral**, reservando el término varianza muestral para la cantidad $\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$

i En estas fórmulas:

Omitimos, si no hay confusión posible, el subíndice n , escribiendo S^2 , $\bar{X} = \sum_{i=1}^n X_i$ y $\bar{X}^2 = \frac{1}{n} \sum_{i=1}^n X_i^2$.

1.16) Esperanza de la varianza muestral

Proposición 1.16.1 Si (X_1, X_2, \dots, X_n) es una muestra aleatoria simple de X con varianza σ_X^2 ,

$$E[S_n^2] = \sigma_X^2.$$

1.17) Distribuciones muestrales de \bar{X} y S^2

💡 Tened en cuenta

- Los resultados anteriores sobre $E[\bar{X}]$ y $\sigma_{\bar{X}}$ son válidos sea cual sea el modelo escogido para la distribución de X .
- Si queremos decir algo más preciso sobre la distribución de \bar{X} (densidad, etc...) necesitamos especificar la distribución de X .
- En el caso en que la variable X siga una distribución normal, el **teorema de Fisher** analiza cómo se comportan los estadísticos anteriores y nos permiten establecer una serie de consecuencias que serán utilizadas posteriormente en los temas de intervalos de confianza y de contrastes de hipótesis.

1.18) Distribución de \bar{X} y S^2 para una m.a.s. de una distribución normal

i Teorema de Fisher

Consideramos una muestra aleatoria simple de una variable aleatoria X con distribución normal $\mathcal{N}(\mu, \sigma^2)$, entonces se verifica:

- 1) \bar{X}_n y S_n^2 son dos variables aleatorias independientes.
- 2) $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$.
- 3) $\frac{(n-1)S_n^2}{\sigma^2} \sim \chi_{n-1}^2$.

1.19) Recordatorio: distribución χ^2 con p grados de libertad

i La distribución χ^2 .

Para $p \in \mathbb{N}^+$, la función de densidad de la distribución ξ^2 es igual a

$$\frac{1}{\Gamma\left(\frac{p}{2}\right) 2^{\frac{p}{2}}} \cdot x^{\frac{p}{2}-1} e^{-\frac{x}{2}}, \text{ si } x > 0,$$

donde Γ denota la función Gamma (Nota: para cualquier real $\alpha > 0$, $\Gamma(\alpha) = \int_0^{+\infty} t^{\alpha-1} e^{-t} dt$).

i Caracterización de la χ^2

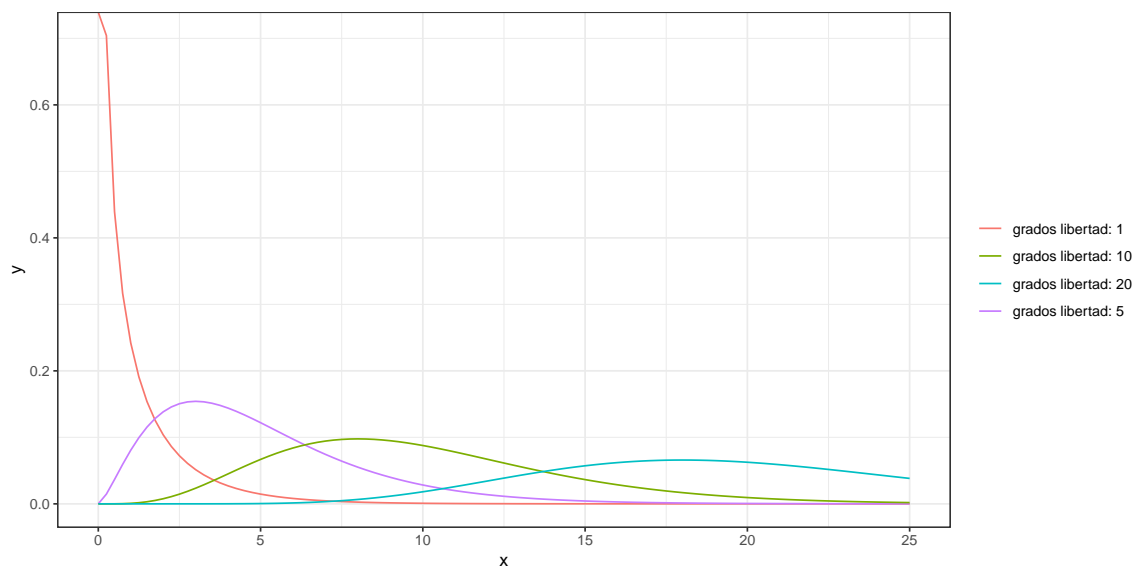
Si Z_1, \dots, Z_p son p variables aleatorias independientes, con $Z_i \sim \mathcal{N}(0, 1)$, entonces la variable aleatoria X definida como

$$X = Z_1^2 + \dots + Z_p^2 = \sum_{i=1}^p Z_i^2$$

tiene una distribución χ^2 con p grados de libertad.

💡 ¿Cómo es su función de densidad?

Depende de los grados de libertad



1.20) Distribución t-Student

i Hemos visto, si X es Normal:

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

Si queremos centrarnos en μ es natural sustituir en ella σ por S_n .

Proposición 1.20.1 Sea (X_1, \dots, X_n) una muestra aleatoria simple de una población $\mathcal{N}(\mu, \sigma^2)$,

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

tiene por densidad

$$f_{n-1}(t) \propto \frac{1}{\left(\frac{1+t^2}{n-1}\right)^{n/2}}, \quad -\infty < t < \infty, \quad (1)$$

La distribución que admite esta densidad se llama **distribución t-Student** con $n-1$ grados de libertad. Escribimos $T \sim t_{n-1}$.

⚠ Su densidad

La función de densidad de una t de Student con k grados de libertad:

$$f_k(t) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)} \frac{1}{\sqrt{k\pi}} \frac{1}{(1+t^2/k)^{\frac{k+1}{2}}}, \quad -\infty < t < \infty,$$

donde Γ denota la función Gamma.

i Caracterización de la t de Student como cociente

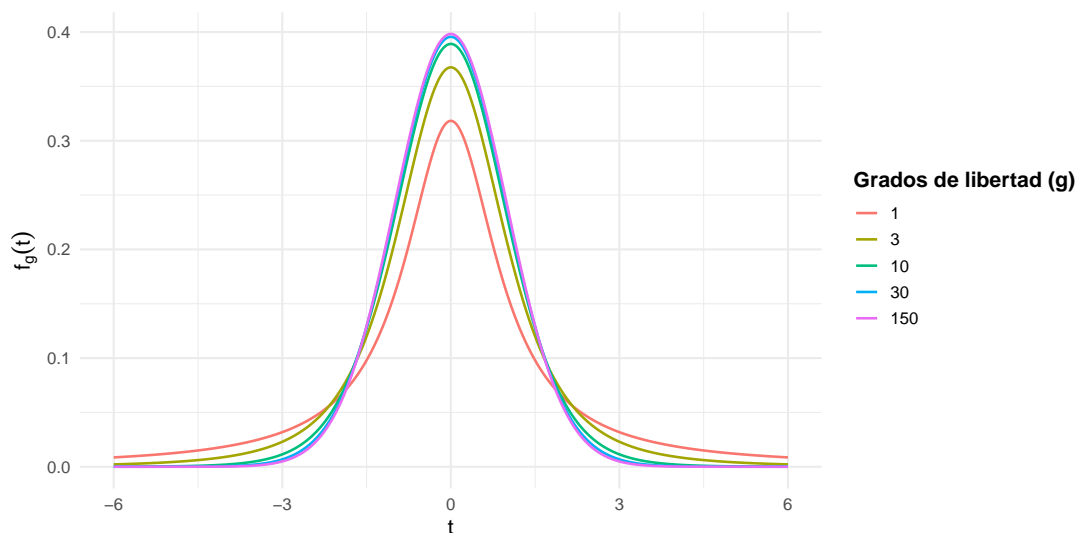
Si Z e Y son dos variables aleatorias independientes, con $Z \sim \mathcal{N}(0, 1)$ e $Y \sim \chi_p^2$, el cociente

$$T = \frac{Z}{\sqrt{Y/p}} \sim t_p,$$

donde t_p denota la t de Student con p grados de libertad.

1.20.1) ¿Cuál es la forma de la densidad de una t -Student?

💡 Tiene colas más pesadas que una normal



Densidad de la distribución t de Student para varios valores de los grados de libertad