Fundamentos de Inferencia Estadística

Francisco Javier Mercader Martínez

Índice

Tema 1: Muestreo y distribuciones muestrales

1.1) Introducción

El contexto

- Tenemos una pregunta acerca de un fenómenos aleatorio.
- Formulamos un modelo para la varaible 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

¿Está la moneda trucada?

- Experimento: tirar la modena. X = resultado obtenido.
  $P(X = +) = p, P(X = c) = 1 - p$

$$¿p = \frac{1}{2}?$$

Sondeo sobre intención de participación en unas elecciones

- Queremos estima 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"). $p = P(X = \text{Si})$.
- ¿Cuánto vale p?
- Censo: aproximadamente 37 000 000. Escogemos aproximadamente 3000 personas.

Determinación de la concentración de un producto

- Quiero determinar la concentración de un producto.
- Formulo el modelo:
  - → Experimento: "llevar a cabo una medición".
  - → 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 estimación 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

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. (⇔ el censo electoral)
    - "1" ⇔ "la persona piensa ir a votar".
    - "0" ⇔ "la persona **no** peinsa ir a votar"
  - → **La tasa de participación "real":** Decido que en mi población el 70% piensa ir a votar → 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.

```
poblacion <- c(rep(1, 25900000), rep(0, 11100000))
set.seed(314159)
p_muestra <- mean(sample(poblacion, 3000, replace = FALSE))
p_muestra
```

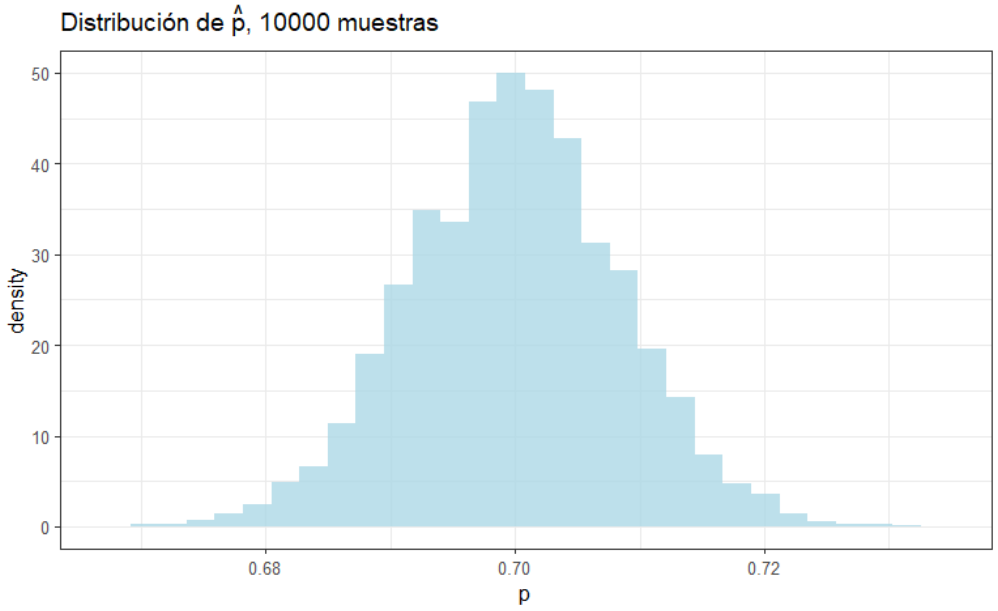
```
## [1] 0.705667
```

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
```

Recogemos los valores obtenidos en un histograma.

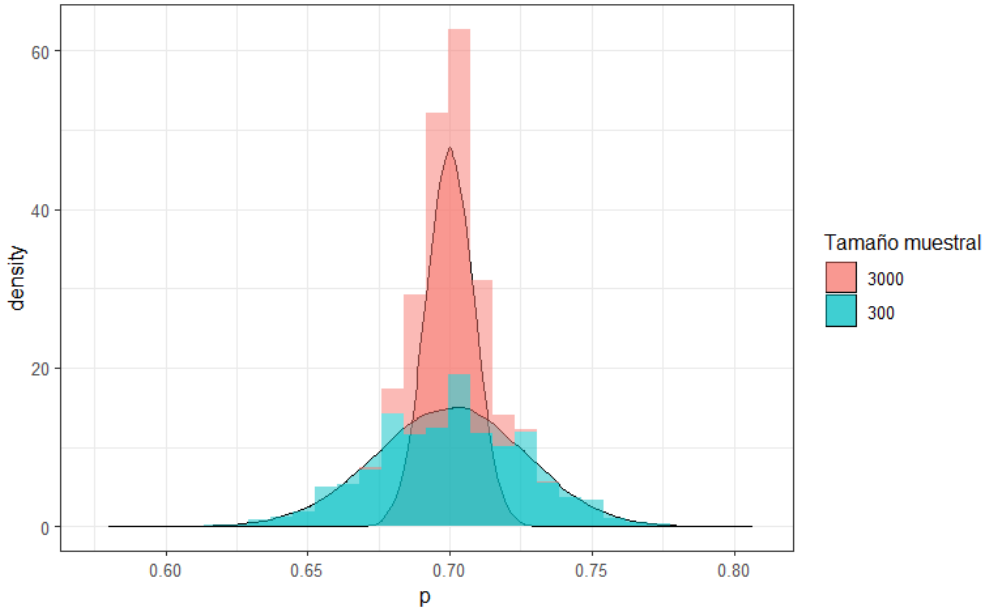


1.5) Realización del experimento: conclusiones

- La enorme mayoría de las muestras de 3000 individuos proporcionan una tasa de partición muy próxima a la de la población.

  → **El riesgo** de cometer un error superior a ±2 puntos, al coger **una** muestra de 3000 individuos es muy pequeño (y asumible...)

- Si nos limitamos a muestras de 300 individuos, ¿qué esperáis?



1.6) En la práctica

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 variables aleatoria: proporción de "1"s en la muestra escogida.
- ¿Cuál es la distribución de valores de $\hat{p}$?

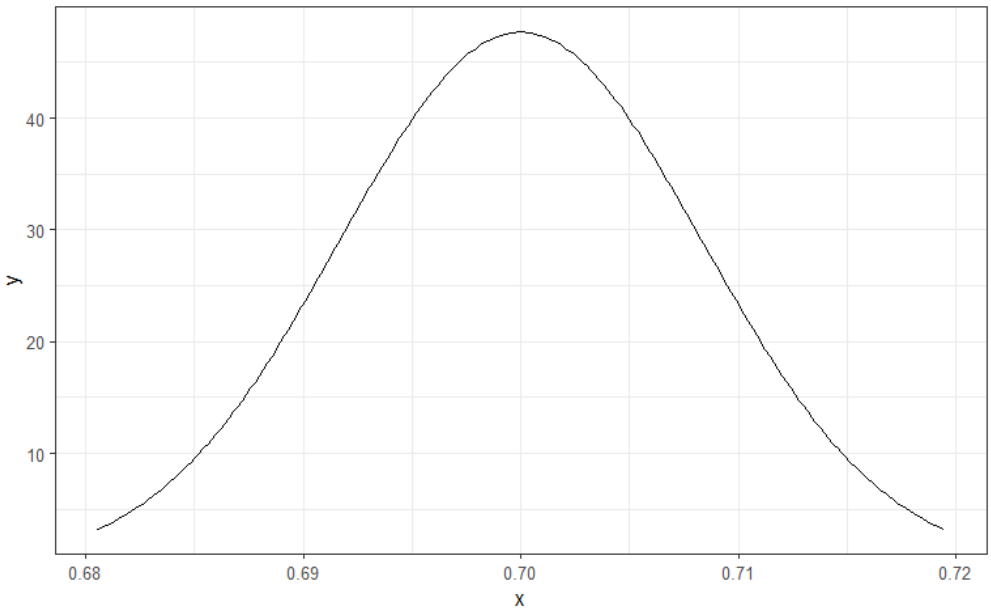
$$\hat{p} \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$$

Es lo qe llamamos la distribución muestral de $\hat{p}$.

1.7) Uso de la distribución muestral

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.



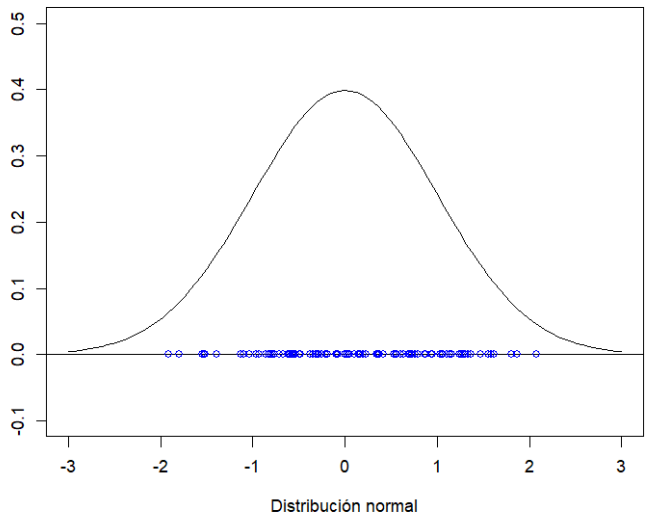
1.8) Antes de extraer una muestra:

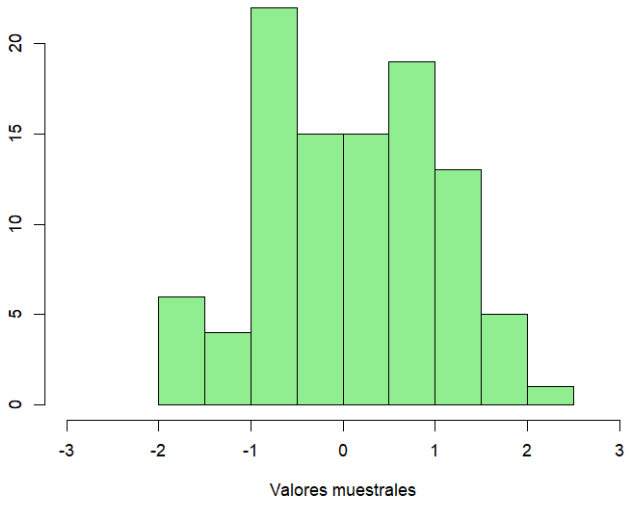
- ¿Es suficiente el tamaño de la muestra para el riesgo asumible y la precisión requerida?
- Una vez extraida la muestra:

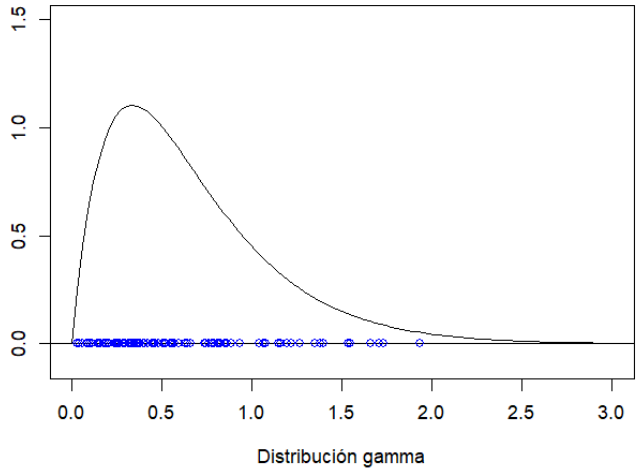
  → ¿Puedo dar un margen de error?

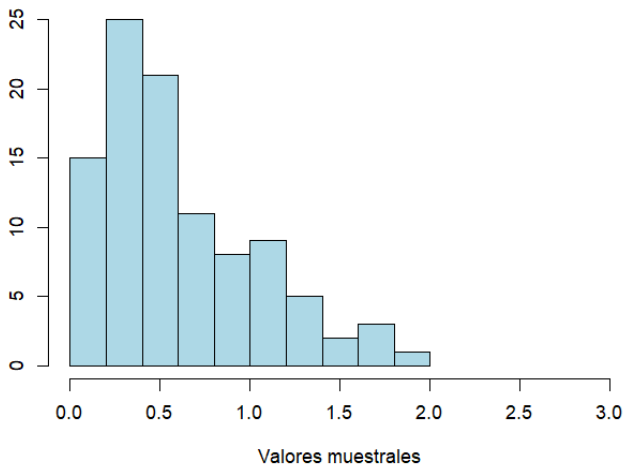
  → ¿Puedo de decidir si p poblacional es, por ejemplo, mayor que un valor dado?

1.9) Otro ejemplo: valores muestrales de una distribución normal









1.10) Un resultado importante

Ley (débil) de los grandes números

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

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

1.11) Algunos términos

Definición

- Sea una variable aleatoria X. Consideramos n variables aleatorias independientes e idénticamente distribuidas $X_1, X_2, \ldots, X_n$, que se distribuyen como X. La variable aleatoria multidimensional $(X_1, X_2, \ldots, X_n)$ es una muestra aleatoria simple (m.a.s) de X.
- Cualquier cantidad calculada a partir de las observaciones de un 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: $\overline{X} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$.
- Desviación típica muestral: $S_X = \sqrt{\dfrac{1}{n+1}\displaystyle\sum_{i=1}^{n}(X_i - \overline{X})^2}$

1.13) La media muestral

Contexto

Estudiamos una variable X cuantitativa.

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

$$x_1, x_2, \ldots, x_n$$

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

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 $\sigma^2 = \text{Var}(X)$.

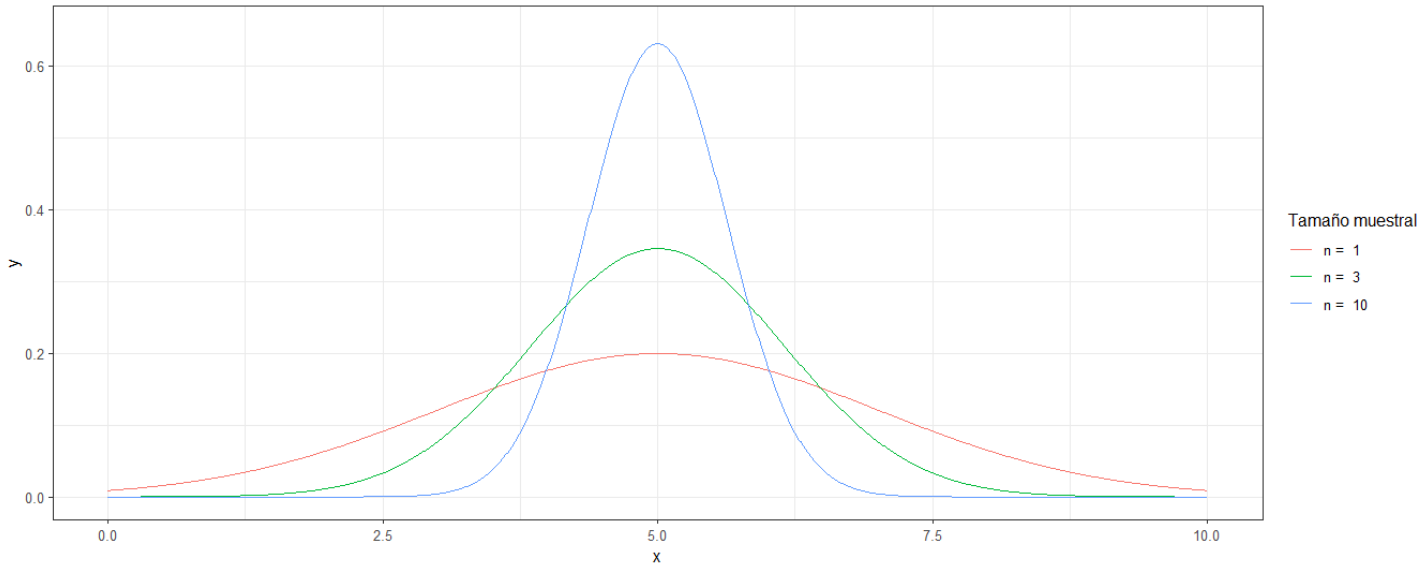
- Tenemos

$$\boxed{E[\overline{X}] = \mu.}$$

  $\rightarrow$ Es decir que el centro de la distribución muestral de $\overline{X}$ coincide con el centro de la distribución X.

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

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

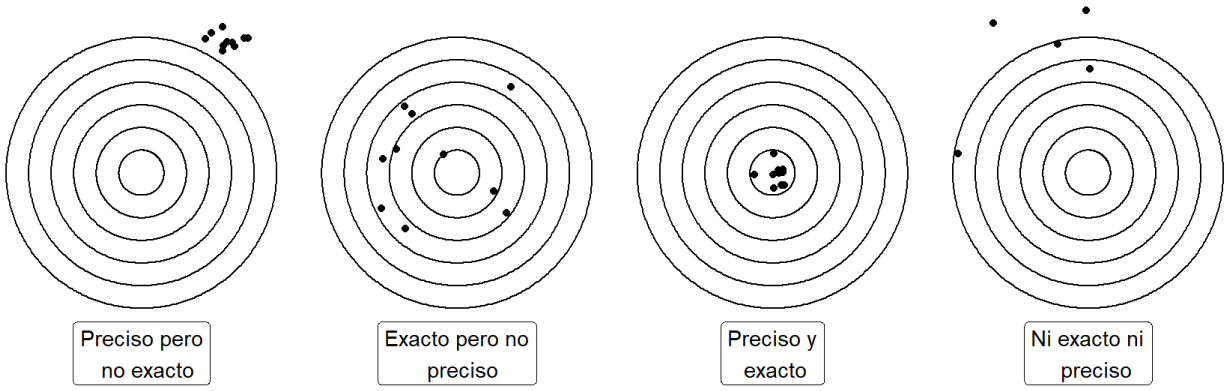


1.14) Consecuencia práctica

Aparato de medición

- Experimento: llevar a cabo una medición con un aparato.
- Variable aleatoria X: "valor propocionado 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 $(X_1, X_2, \ldots, X_n)$ es una muestra aleatoria simple de X, definimos la varianza muestral S_n^2 como

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2.$$

Fórmula alternativa para S_n^2:

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

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

1.15.1) Dos apuntes

En algunos textos en castellano:

Se suele llama S_n^2 cuasi-varianza muestra, reservando el término varianza muestral para la cantidad $\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2$.

En estas fórmulas:

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

1.16) Esperanza de la varianza muestral

Proposición

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

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

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

Tened en cuenta

- Los resultados anteriores sobre $E[\overline{X}]$ y $\sigma_{\overline{X}}$ son válidos sea cual sea el modelo escogido para la distribución de X.
- Si queremos decir algo más preciso sobre la distribución de $\overline{X}$ (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 constrastes de hipótesis.

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

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

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

La distribución χ^2.

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

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

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

Caracterización de la χ^2

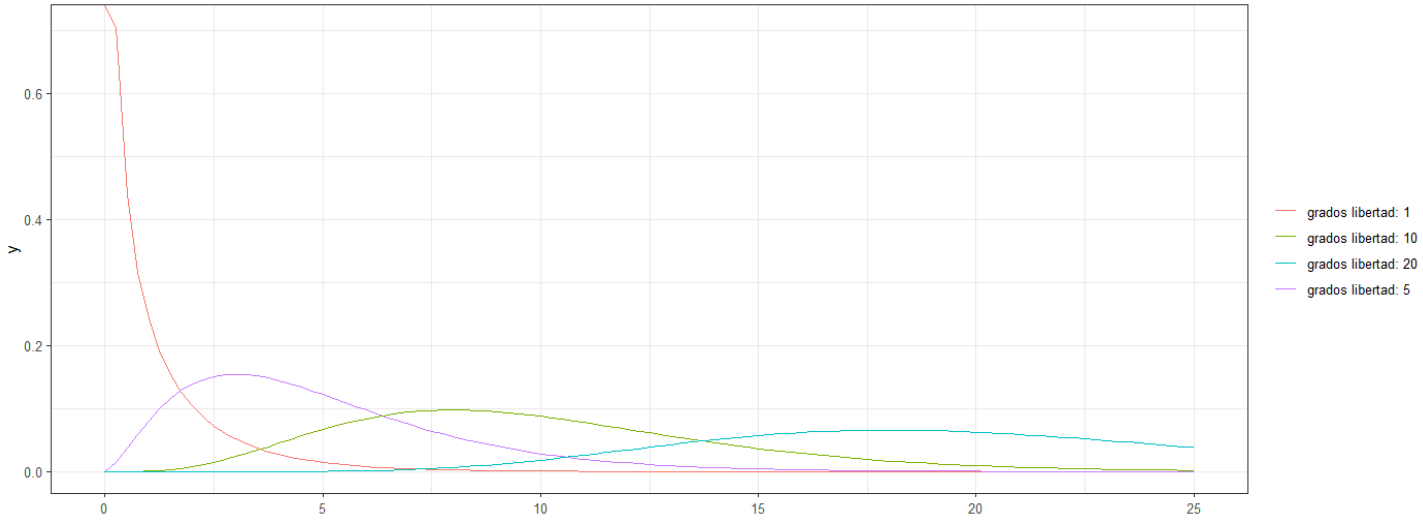
Si $Z_1, \ldots, Z_p$ son p variables aleatorias independientes, con $Z_i \sim \mathcal{N}(0,1)$, entonces la variable aleatoria X definida como

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

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

Hemos visto, si X es Normal:

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

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

Proposición

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

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

tiene por densidad

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

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 un t-Student con k grados de libertad:

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

donde Γ denota la función Gamma.

Caracterización de la t-Student como cociente

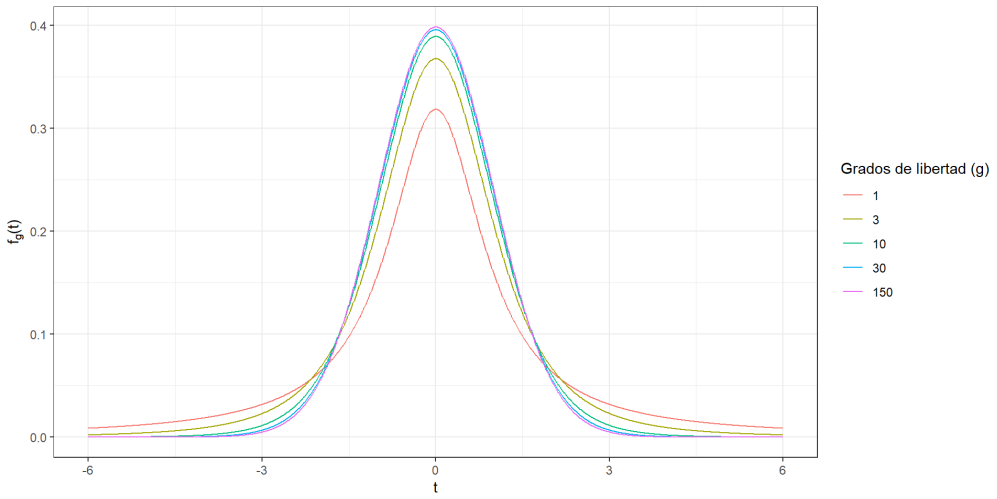
Si Z e Y son dos variables aleatorias independientes, con $Z \sim \mathcal{N}(0,1)$ e $Y \sim \chi^2_p$, el cociente

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

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

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

Tiene colas más pesadas que una normal



1.21) Distribución F de Snedecor para el cociente de varianzas

Proposición

Consideremos U_1 y U_2 dos variables aleatorias independientes con distribución χ^2 con p_1 y p_2 grados de libertad, respectivamente.

El cociente $F = \dfrac{\frac{U_1}{p_1}}{\frac{U_2}{p_2}}$ admite la densidad

$$f_F(x) = \frac{\Gamma\left(\frac{p_1+p_2}{2}\right)}{\Gamma(p_1)\Gamma(p_2)} \left(\frac{p_1}{p_2}\right)^{p_1} \frac{x^{\frac{p_1}{2}-1}}{\left(1+\frac{p_1}{p_2}x\right)^{\frac{p_1+p_2}{2}}}.$$

Esta distribución se llama F de Snedecor p_1 y p_2 grados de libertad y escribimos $F \sim F_{p_1,p_2}$.

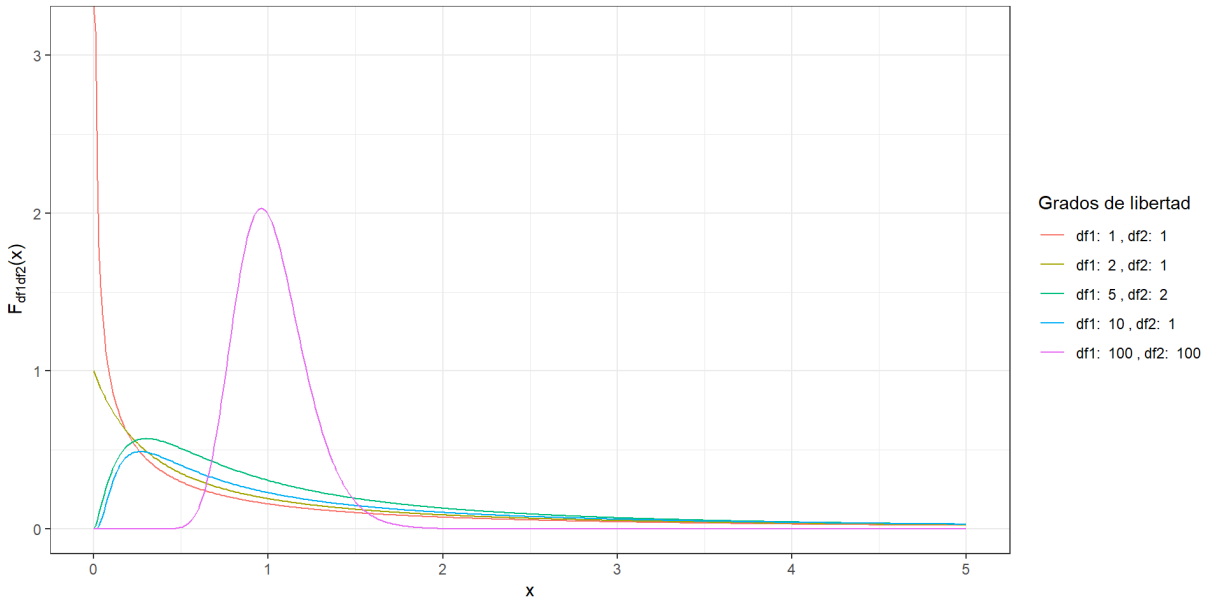
Consecuencia

Consideremos X e Y variables aleatorias normales independientes con varianzas σ_X^2 y σ_Y^2, así como $X_1,\ldots,X_{n_x}$ e $Y_1,\ldots,Y_{n_Y}$ dos muestras aleatorias simples de X e Y, respectivamente. Deducimos que

$$\frac{\frac{S_X^2}{\sigma_X^2}}{\frac{S_Y^2}{\sigma_Y^2}} \sim F_{n_X-1,n_Y-1}.$$

- ¿Cuál es la forma de la densidad de una F de Snedecor?

Depende mucho de los grados de libertad



1.22) Si la distribución de X no es Normal

No podemos decir nada en general, excepto si n es grande...

Teorema Central del Límite

Si n es *"suficientemente"* grande, se puede aproximar la distribución de $\overline{X}$ por una Normal con media μ y varianza $\frac{\sigma^2}{n}$:

$$\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \text{ aproximadamente.}$$

Formulación matemática

El resultado anterior se taduce por una convergencia de la sucesión de las variables aleatorias $\left(\overline{X}_n\right)_n$ en distribución cuando $n \to \infty$.

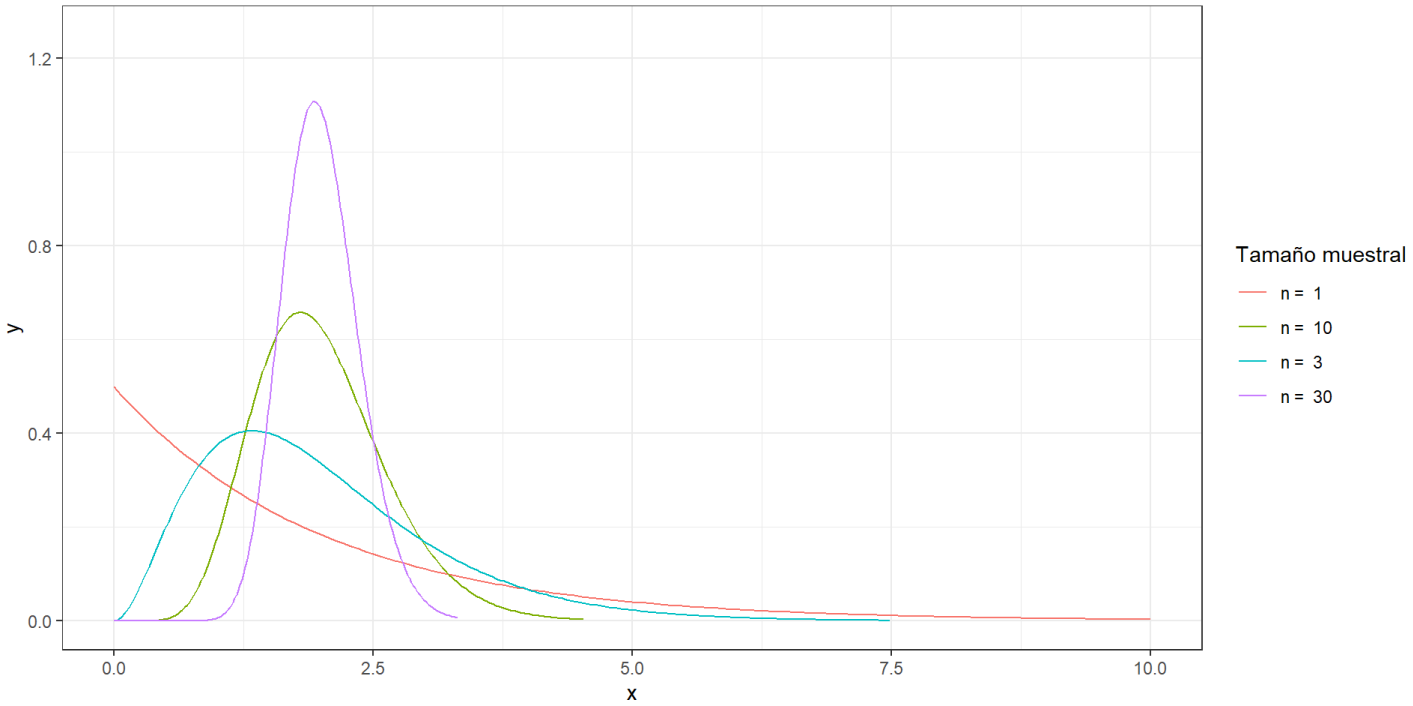
- ¿Cuándo considerar que n es grande?

Depende de la forma de la distribución de X:

- Si X casi Normal: n pequeño es suficiente.
- Si X es muy asimétrico: n mucho más grande necesario.

En general, se suele considerar $n \geq 30$ suficiente...

Ilustración, X inicial $\sim \text{Exp}(\lambda = 0.5)$, $\overline{X}$ **con** $n = 3, 10$ **y** $n = 30$


1.2
0.8
0.4
0.0
y
0.0
2.5
5.0
7.5
10.0
x
Tamaño muestral
n = 1
n = 10
n = 3
n = 30