Distribuciones Muestales

Sergio Nava

octubre de 2022

1 Distribuciones Muestrales

Section 1

Distribuciones Muestrales

Muestreo

- El muestreo es una herramienta de la investigación científica.
- Su función básica es determinar que parte de una realidad en estudio (población o universo) debe examinarse con la finalidad de hacer inferencias sobre dicha población.

Revisión de conceptos

- La información de los estudios de muestreo es parte de nuestra vida diaria, casi en su totalidad. Tal información determina el rumbo que deberán tomar algunas políticas gubernamentales como, por ejemplo, la promoción de programas sociales o el control de la economía
- Las encuestas de opinión son la base de muchas de las noticias proporcionadas en los medios. Los estudios de rating televisivo determinan cuales son los programas que permanecerán al aire en el futuro.
- No se diga los estudios de preferencias electorales, para definir estrategias por parte de los partidos políticos.
- Las investigaciones de mercado indicaran cuales productos y con que características son los preferidos de los consumidores
- Por otro lado, están los estudios de muestreo en las ciencias biológicas, geológicas, del medio ambiente, marítimas entre otras.
- Muestreo de Aceptación (Industrial)

- Aún cuando la terminología de las ciencias sociales difiere de las ciencias exactas, los científicos sociales conducen estudios de muestreo y los científicos de las áreas físicas realizan en su mayoria experimentos, ambos tienen el propósito de captar información en torno a los fenómenos naturales.
- Sin embargo, esas diferencias existen en el campo de la ciencia, debido a la naturaleza de las poblaciones y a la manera en que una muestra puede ser extraída. Por ejemplo, poblaciones de votantes, de cuentas financieras, o de animales de una especie particular pueden contener un número relativamente pequeño de elementos (finito).
- En contraste, la población conceptual de respuestas generadas por la medición de la producción de un proceso químico, es muy grande (infinito). Las limitaciones del procedimiento de muestreo también varían de un área de la ciencia a otra.
- El muestreo en las ciencias biológicas y físicas, puede frecuentemente ser realizado bajo condiciones experimentales controladas. Tal control es frecuentemente imposible en las ciencias sociales, negocios, y administración de recursos naturales (observación).

Un ejemplo: Población y muestra

¿Cómo realizar un inventario?

1. Censo: es un conteo exhaustivo de los individuos o elementos de la población bajo estudio.

Desventajas:

- Costos elevados.
- Estático
- Requiere mucho tiempo

2. Muestreo: una parte representativa del recurso.

Ventajas:

- Reduce costos.
- Puede ser dinámico
- Reduce tiempos.

Conceptos de Población y Muestra

- Se ha manejado que la estadística moderna es la teoría de la información, cuyo objetivo es la *inferencia.* Nuestro interés se centra en un grupo de mediciones que existen o pueden ser generadas, una población. El medio de la inferencia es la muestra, la cual es un subgrupo de mediciones seleccionadas de la población.
- Deseamos entonces realizar inferencias sobre la población basándonos en las características que observamos en la muestra, o equivalentemente, en la información contenida en la muestra.

N elementos de la población

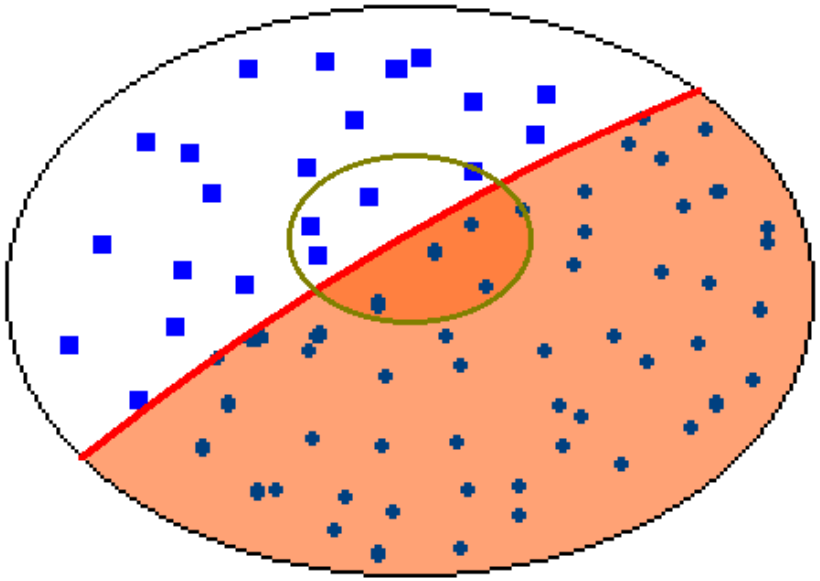


Figure 1: Población y muestra

n elementos de la muestra

Error de muestreo

- El error que se comete debido al hecho de que se obtienen conclusiones sobre cierta realidad a partir de la observación de sólo una parte de ella, se denomina **error de muestreo**.
- Obtener una muestra adecuada, significa lograr una versión simplificada de la población, que reproduzca de algún modo sus rasgos y características básicas o de interés.

Terminología

- **Elemento** es un objeto o persona en el cual se toman las mediciones.
- **Población objetivo**: conjunto de individuos de los que se quiere obtener información.
- **Unidades de muestro**: el conjunto de elementos no traslapados de la población que cubren a la población completa. Todo miembro de la población pertenecerá a una y sólo una unidad de muestreo.
- **Unidades de análisis**: objeto o individuo del que hay que obtener la información.
- **Marco muestral**: lista de unidades o elementos de muestreo.
- **Muestra**: conjunto de unidades o elementos de análisis seleccionadas de un marco o varios marcos.

Terminología

- **Muestreo probabilístico**. El planteamiento clásico del problema de estimación estadística requiere que la aleatoriedad esté comprendida en el diseño de muestreo para así poder evaluar probabilísticamente, las propiedades de los estimadores. Al diseño de muestreo que plantea la selección, de unidades de muestreo, basada en la aleatoriedad se le llama *muestreo probabilístico*.

Terminología

- **Límite para el error de estimación**. Si θ es la característica poblacional de interés y $\hat{\theta}$ es un estimador (basándose en la información de la muestra) de θ, debemos especificar un límite para el error de estimación; esto es, debemos especificar que θ y $\hat{\theta}$ difieran en valor absoluto a lo más en cierto valor B. Simbólicamente,

$$\text{error de estimación} = |\theta - \hat{\theta}| < B$$

- θ puede ser cualquier característica de la población (el promedio, el total, un porcentaje, el valor mediano, el valor mínimo, etcétera) Se le llama **parámetro**.
- $\hat{\theta}$ es el **estadístico** obtenido a partir de la información de la muestra. En algunas veces llamado estadístico de prueba. (el promedio de la muestra, el total de la muestra, el mínimo de la muestra, la mediana de la muestra, etcétera)

Parámetro poblacional vs Estadístico muestral

- **Parámetro:** Es una cantidad numérica calculada sobre una población
  - La altura media de los individuos de un país
  - La idea es resumir toda la información que hay en la población en unos pocos números (parámetros).
- **Estadístico:** Es una cantidad numérica calculada sobre una muestra de la población
  - La altura media de los que estamos en este aula.
    - ¿Somos una muestra de la población? ¿representativa?
  - Si un estadístico se usa para aproximar un parámetro también se le suele llamar estimador.[a]

[a] Normalmente nos interesa conocer un parámetro, pero por la dificultad que conlleva estudiar a *TODA* la población, calculamos un estimador sobre una muestra y "confiamos" en que sean próximos. Más adelante veremos como elegir muestras para que el error sea "confiablemente" pequeño

	Población	Muestra
	Parámetro	Estadístico
Media	μ	$\bar{x}$
Proporción	P	p
Máximo	max	max
Mediana	$\tilde{x}$	
Varianza	σ^2	s^2
Total	T	$\hat{T}$
	θ	$\hat{\theta}$

- También debemos definir una probabilidad, $(1-\alpha)$ que especifique la fracción de veces en muestreo repetido, que requeriremos que el error de estimación sea menor que B. Esto es

$$P[\text{error de estimación} < B] = 1 - \alpha$$

- **Muestreo no probabilístico**. El muestreo no probabilístico no involucra ningún elemento aleatorio en el proceso de selección.

Definición: Si $X_1, X_2, ..., X_n$ son **variables aleatorias** independientes e idénticamente distribuidas, decimos que constituyen una **muestra aleatoria** de la población **infinita** dada por su distribución común.

Si S es un espacio muestral con una medida de probabilidad y X es una función con valor real definida con respecto a los elementos de S, entonces X se denomina **Variable Aleatoria**.

Muestreo Aleatorio Simple (Población Finita)

Unas muestra aleatoria simple de tamaño n, de una población finita de tamaño N, es una muestra seleccionada de tal manera que cada una de las muestras posibles de tamaño n tenga la misma probabilidad de ser seleccionada.

Distribución de Muestreo

¿Qué es una distribución muestral?

La distribución muestral de un estadístico de prueba proporciona

1. una lista de todos los valores que puede tomar dicho estadístico y
2. la probabilidad de obtener cada valor, suponiendo que éste es producto sólo del azar.

Distribución muestral de la media

La distribución muestral de la media proporciona todos los valores que puede tomar la media, junto con la probabilidad de obtener cada valor si el muestreo es aleatorio a partir de la población hipotética.

La media muestral posee las siguientes características:

1. $\mu_{\bar{x}}=$ es la media de la distribución muestral de la media.

   $s_{\bar{x}}$ = es la desviación estándar de la distribución muestral de la media
2. La media muestral es igual a la media poblacional, $\mu_{\bar{x}} = \mu$.
3. La media muestral tiene una desviación estándar igual a la desviación estándar poblacional de datos crudos, dividida entre la raíz del número de datos. Es decir: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
4. Presenta una forma de campana.

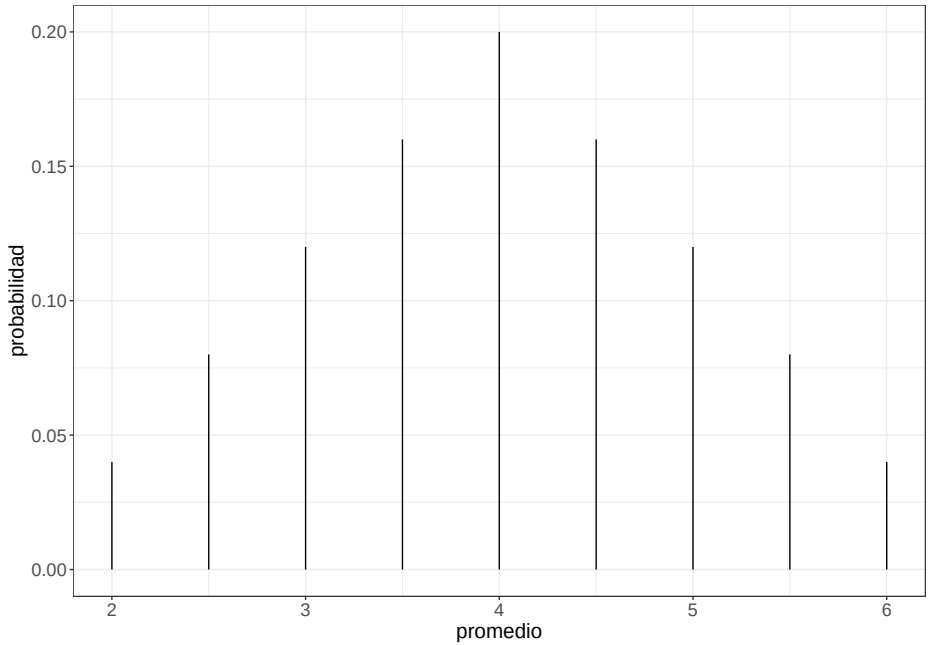
A pesar que la demostración de la distribución muestral de la media va más allá de los alcances del curso, podemos hacer un ejemplo para la mejor comprensión de la distribución muestral de la media.

Supongamos una población de solo cinco elementos $2, 3, 4, 5$ y 6. La media μ de la población es $\mu = 4.0$ y la desviación estándar de la población es $\sigma = 1.41$.

Ahora queremos deducir la distribución muestral de la media para muestras de tamaño 2 de la población. Extraemos (con reemplazo) todas las distintas muestras de tamaño $n = 2$. Y observamos cual es el valor de $\bar{x}$ y su probabilidad.

Var1	Var2	promedio	muestra
2	2	2.0	1
3	2	2.5	2
4	2	3.0	3
5	2	3.5	4
6	2	4.0	5
2	3	2.5	6
3	3	3.0	7
4	3	3.5	8
5	3	4.0	9
6	3	4.5	10
2	4	3.0	11
3	4	3.5	12
4	4	4.0	13
5	4	4.5	14
6	4	5.0	15
2	5	3.5	16
3	5	4.0	17
4	5	4.5	18
5	5	5.0	19
6	5	5.5	20
2	6	4.0	21
3	6	4.5	22
4	6	5.0	23
5	6	5.5	24
6	6	6.0	25

promedio	conteo	p
2.0	1	0.04
2.5	2	0.08
3.0	3	0.12
3.5	4	0.16
4.0	5	0.20
4.5	4	0.16
5.0	3	0.12
5.5	2	0.08
6.0	1	0.04



Var1	Var2	promedio	muestra
2	2	2.0	1
3	2	2.5	2
4	2	3.0	3
5	2	3.5	4
6	2	4.0	5
2	3	2.5	6
3	3	3.0	7
4	3	3.5	8
5	3	4.0	9
6	3	4.5	10
2	4	3.0	11
3	4	3.5	12
4	4	4.0	13
5	4	4.5	14
6	4	5.0	15
2	5	3.5	16
3	5	4.0	17
4	5	4.5	18
5	5	5.0	19
6	5	5.5	20
2	6	4.0	21
3	6	4.5	22
4	6	5.0	23
5	6	5.5	24
6	6	6.0	25

Media de la población

$$\mu = \frac{\sum X}{N} = 4.0$$

Media de la medias muestrales

$$\mu_{\bar{x}} = \frac{\sum \bar{x}}{m} = \frac{100}{25} = 4.0$$

Así, $\mu_{\bar{x}} = \mu$. También del resultado podemos verificar que:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1.41}{\sqrt{2}} = 1.0$$

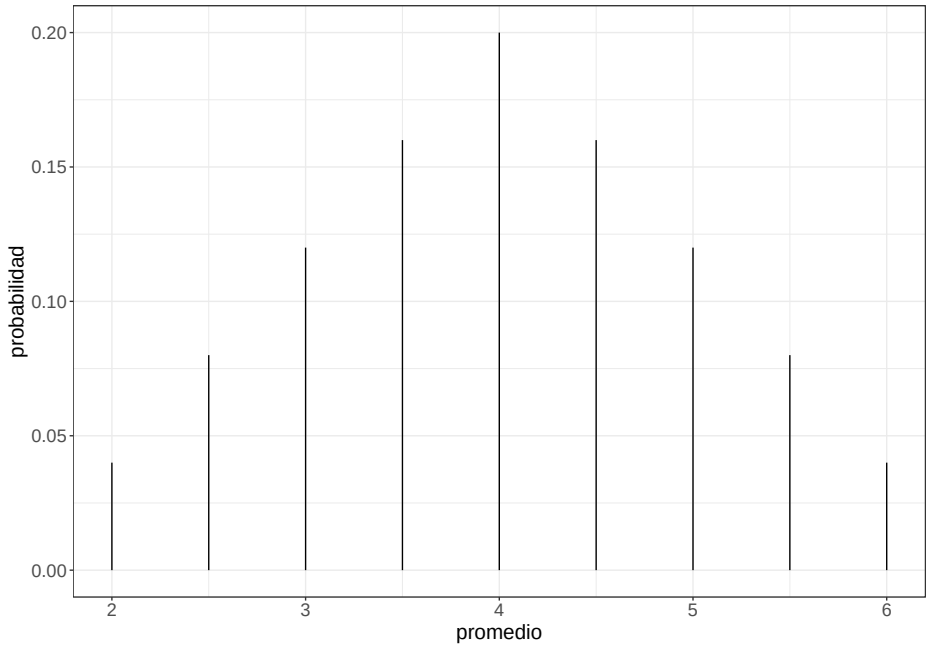
Pero podemos calcular directamente:

$$\sigma_{\bar{x}} = \sqrt{\frac{\sum(\bar{x} - \mu_{\bar{x}})^2}{m}} =$$

$$= \sqrt{\frac{(2-4)^2 + \cdots (6-4)^2}{25}} = 1.0$$

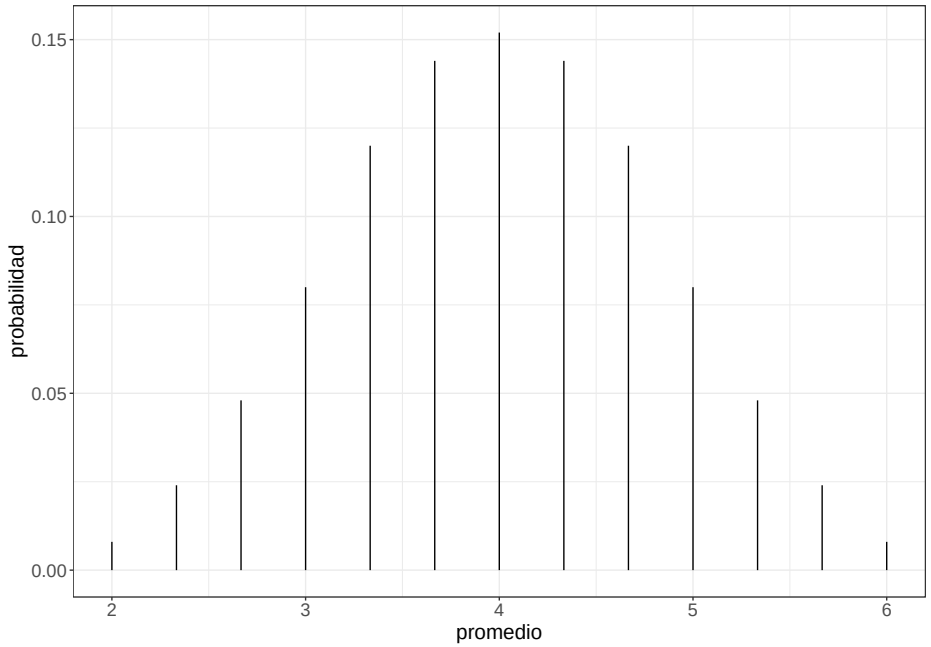
Distribución muestral de la media para N=5, con $\mu = 4$, $\sigma = \sqrt{2}$ y $n = 2$

promedio	conteo	p
2.0	1	0.04
2.5	2	0.08
3.0	3	0.12
3.5	4	0.16
4.0	5	0.20
4.5	4	0.16
5.0	3	0.12
5.5	2	0.08
6.0	1	0.04

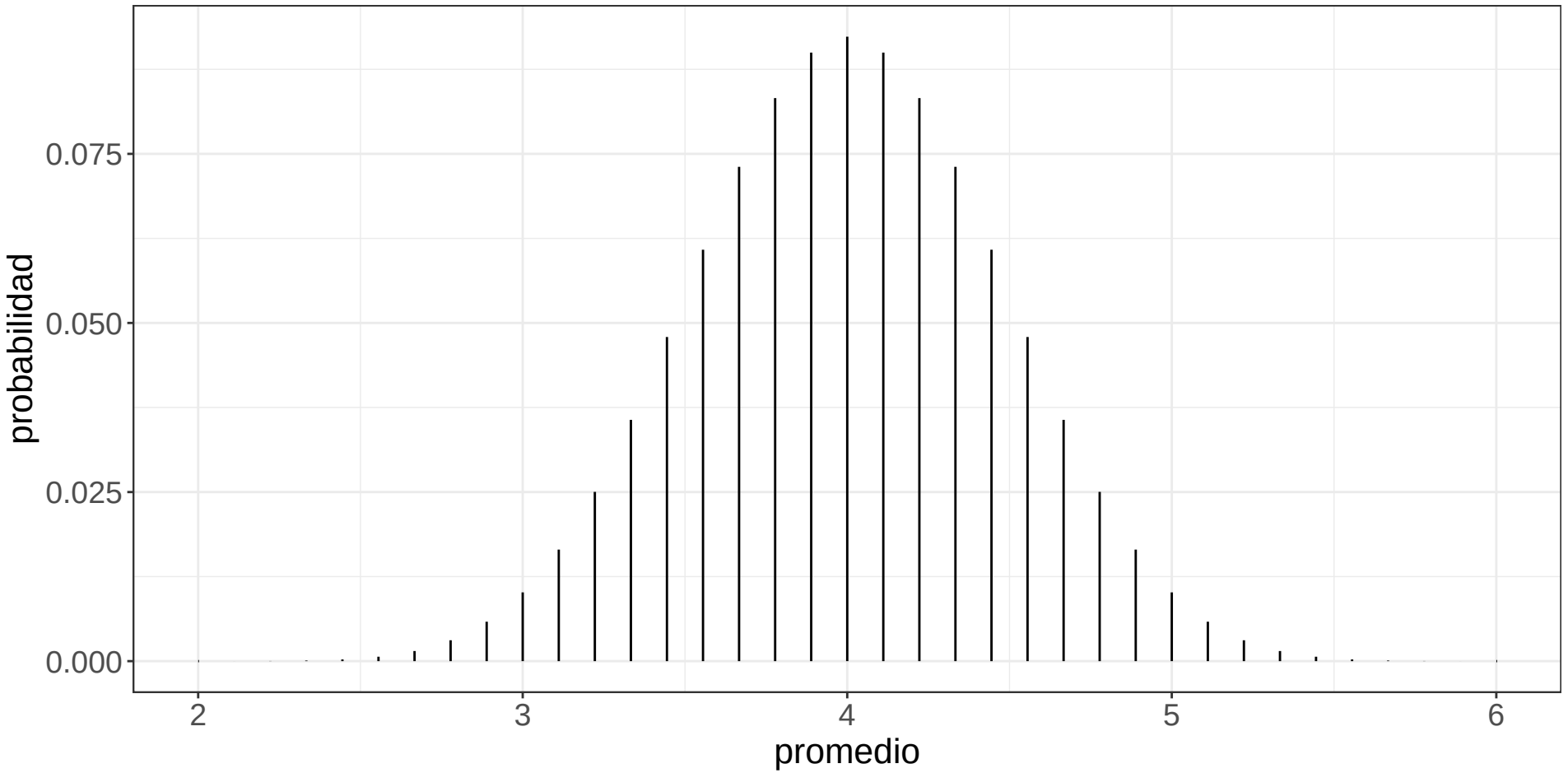


Distribución muestral de la media para N=5, con $\mu = 4$, $\sigma = \sqrt{2}$ y $n = 3$

promedio	conteo	p
2.000	1	0.008
2.333	3	0.024
2.667	6	0.048
3.000	10	0.080
3.333	15	0.120
3.667	18	0.144
4.000	19	0.152
4.333	18	0.144
4.667	15	0.120
5.000	10	0.080
5.333	6	0.048
5.667	3	0.024
6.000	1	0.008



Distribución muestral de la media para N=5, con $\mu = 4$, $\sigma = \sqrt{2}$ y $n = 9$



Ver el archivo **distr_muestral_pequena.R** en https://rstudio.cloud/content/4731243

Teorema del Límite Central

Central Limit Theorem

Teorema: Si $X_1, X_2, \ldots, X_n$ constituyen una muestra aleatoria de una población infinita que tiene la media μ y la varianza σ^2, entonces la distribución límite de:

$$\bar{x} \sim N\left(\mu_{\bar{x}} = \mu, \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\right)$$

cuando $n \to \infty$. Si definimos

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

entonces $Z \sim N(0, 1)$ (la distribución normal estándar).

Una forma sencilla de expresar el teorema del límite central es: "la suma (o promedio) de n variables aleatorias independientes e idénticamente distribuidas $(i.i.d.)$, sigue una distribución límite normal con media $n\mu$ (ó μ) y varianza σ^2 (σ^2/n)".

Ejemplo

Una maquina vendedora de refrescos está programada para que la cantidad de refresco que se sirva sea una variable aleatoria con una media de 200 mililitros y una desviación estándar de 15 mililitros. ¿cuál es la probabilidad de que la cantidad de refresco promedio (media) servida en una muestra tomada al azar de 36, sea cuando menos 204 mililitros.

Solución

Según el TLC, la distribución de $\bar{x}$ tiene la media $\mu_{\bar{x}} = 200$ y desviación estándar $s_{\bar{x}} = 15/\sqrt{36} = 2.5$, y tiene una distribución que es aproximadamente normal. Como $z = (204 - 200)/2.5 = 1.6$, podemos calcular la probabilidad $P(\bar{x} \geq 204) = P(z \geq 1.6) = 0.0548$.

Para que se alcance una distribución parecida a la normal en el conjunto de posibles promedios muestrales se requiere que n sea grande.

Sin embargo, la rapidez de acercamiento a la normal (velocidad de convergencia) también depende de la forma de la distribución de la variable en la población.

Ver la siguiente liga de de una aplicación hecha en shinny donde se puede ver el histograma de promedios, donde sepuede modificar el número de simulaciones y el tamaño de la muestra n de cuatro distribuciones distintas.

https://s3rgionava.shinyapps.io/histograma_de_medias/

Distribución de la media (población finita)

Si $\bar{x}$ es la media de una muestra aleatoria de tamaño n tomada de una población finita de tamaño N con media μ y la varianza σ^2, entonces:

$$E(\bar{x}) = \mu \text{ y } Var(\bar{x}) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$$

Distribución muestral de proporciones.

Distribución Binomial

Definición: Si $X_1, X_2, \ldots, X_n$ son variables aleatorias independientes e idénticamente distribuidas $(i.i.d.)$ que solo toman valores de $X_i = 1$ ó $X_i = 0$, dependiendo si poseen o no la característica de interés respectivamente, decimos que constituyen una **muestra aleatoria de un experimento binomial** de la población infinita dada por su distribución común.

Definición: Si $X_1, X_2, \ldots, X_n$ constituyen una muestra aleatoria de un experimento binomial, entonces

$$p = \frac{\sum_{i=1}^{n} x_i}{n}$$

se denomina **proporción de la muestra** y

$$np(1-p)$$

es la **varianza de la muestra**.

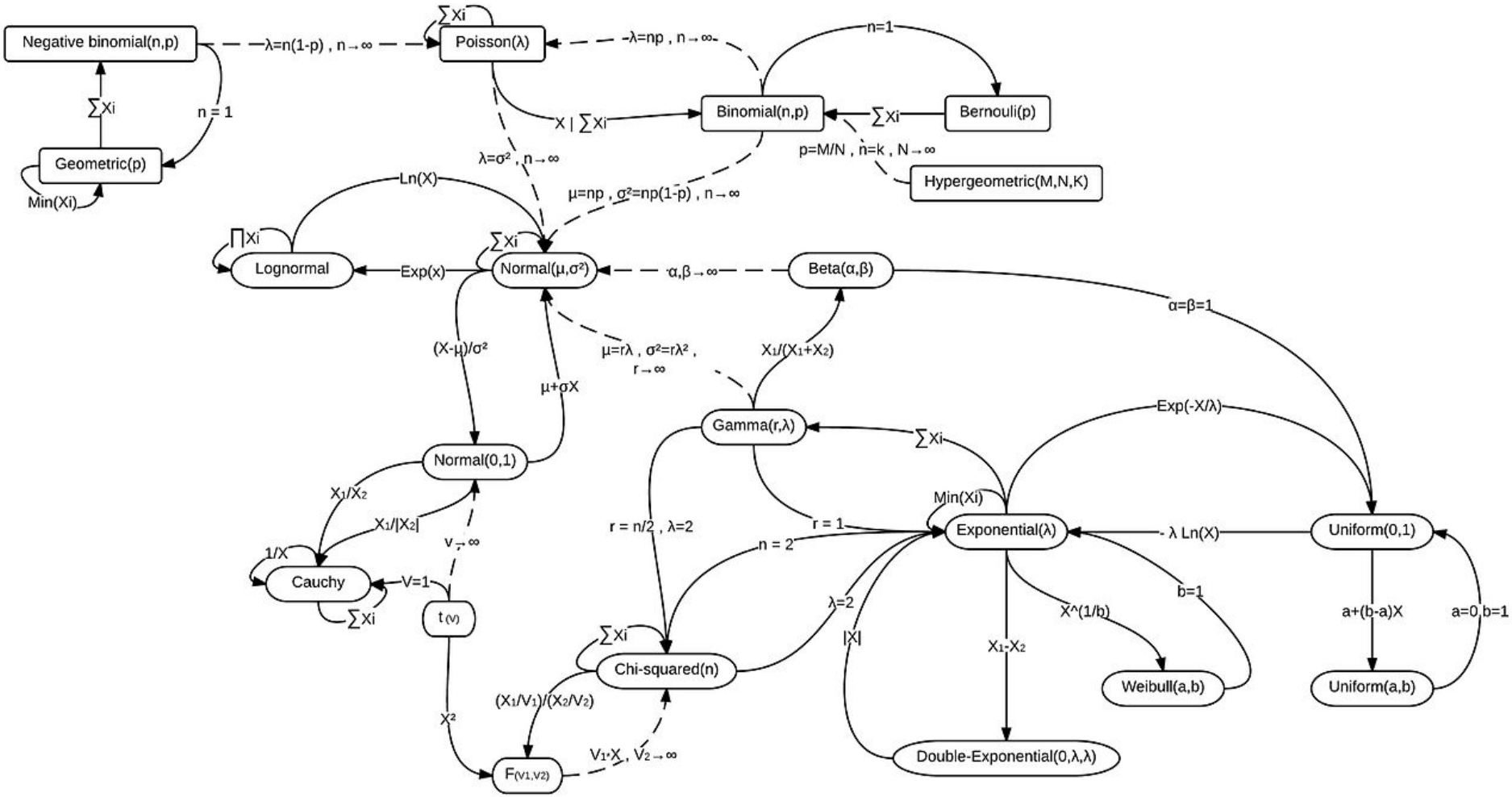


Figure 2: Relaciones entre distribuciones de probabilidad

Para verlo a detalle hacer clic aquí

Para un tamaño de muestra suficientemente grande

Teorema: Si $X_1, X_2, \ldots, X_n$ constituyen una muestra aleatoria de una población infinita donde X_i constituye un experimento Bernoulli, tal que que P es la proporción de la población con la característica de interés, entonces se cumple que:

$$E(p) = P \text{ y } Var(p) = \frac{p(1-p)}{n}$$

Corolario: De aquí y del TLC se tiene que la distribución límite de

$$z = \frac{p - P}{\sqrt{\frac{p(1-p)}{n}}}$$

cuando $n \to \infty$ es normal estándar.

Ejemplo

La proporción de familias de la ciudad de Aguascalientes, que son dueñas (no arrendatarias) de sus casas es de 0.70. Si al azar se entrevistan a 84 familias de esta ciudad y sus respectivas respuestas (a la pregunta de si son dueñas o no de su casa) se consideran valores de variables aleatorias independientes que tienen distribución de Bernoulli idénticas con el parámetro $P = 0.70$, ¿Con qué probabilidad podemos afirmar que el valor que se obtenga de la muestra p será menor que 0.64

Respuesta

$P = 0.70$, $p = 0.64$, $n = 84$, sustituyendo

$$z = \frac{p - P}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0.64 - 0.70}{\sqrt{\frac{0.64(0.36)}{84}}} = -1.1456$$

$$p(x < 64) = p(z < -1.1456) = 0.1259$$

Aproximación para proporciones

También se puede ver este resultado como dada una muestra aleatoria $X_1, X_2, \ldots, X_n$ de variables aleatorias Bernoulli, con $x = \sum_{i=1}^{n} x_i$ el número de éxitos observados, en n intentos igualmente probables, entonces la distribución límite de:

$$z = \frac{x - np}{\sqrt{np(1-p)}}$$

con $p = x/n$, para $n \to \infty$, z tiene una distribución límite normal estándar.

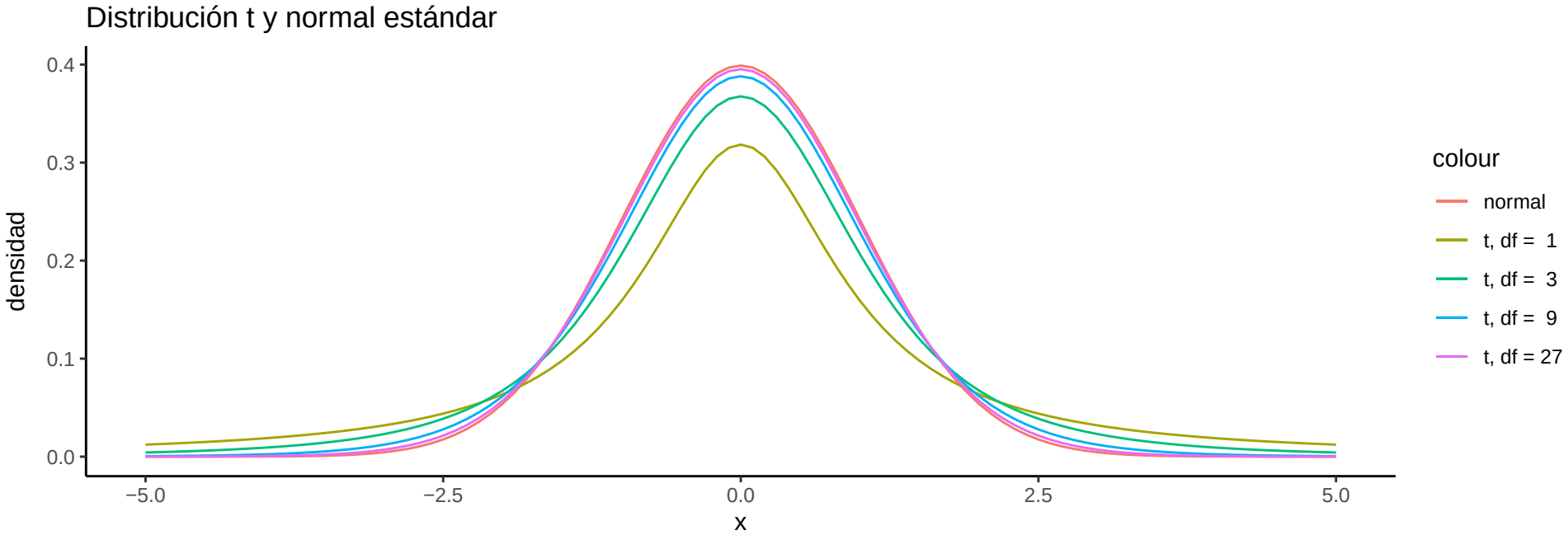
Distribución t-Student (William S. Gosset)

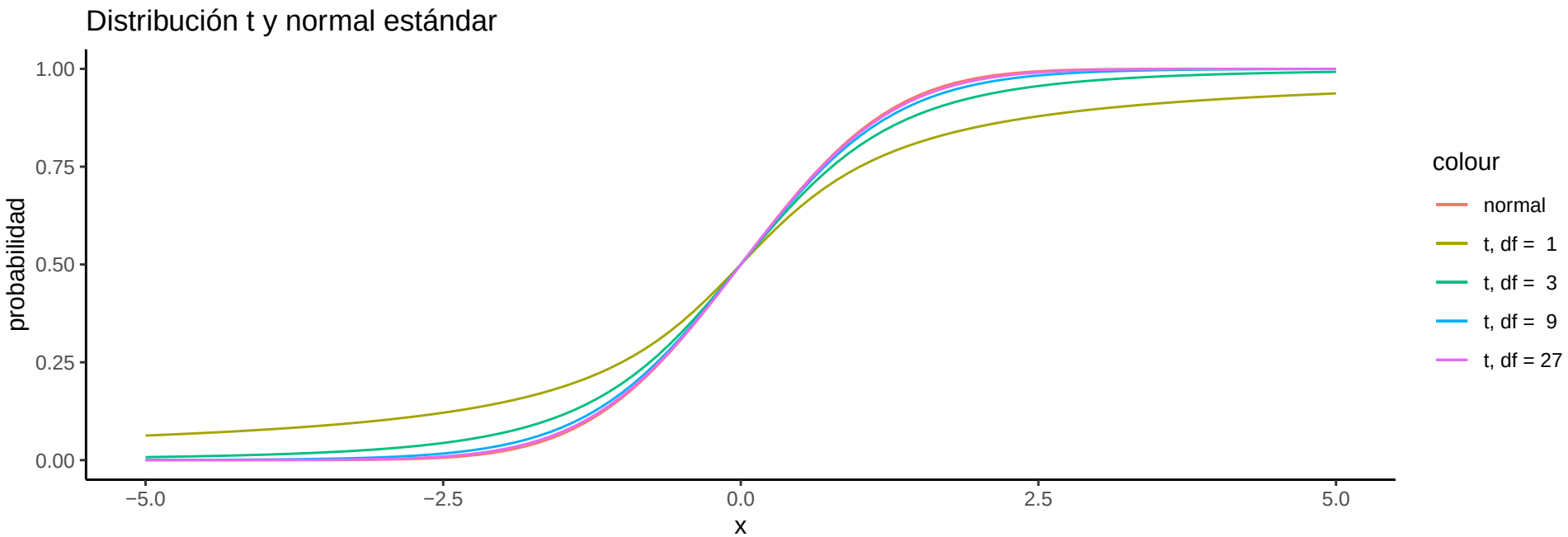
Si $\bar{x}$ y s^2 son la media y la varianza de una muestra aleatoria de tamaño n tomada de una población **normal** con media μ , entonces

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

tiene una distribución *t-student* con $n - 1$ grados de libertad.
Para usar la distribución normal es necesario conocer el valor de la desviación estandar poblacional σ. Como es más común el desconocimiento, entonces se estima σ a través de s (desviación estándar muestral) y se usa la distribución t.

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$


Distribución t y normal estándar
densidad
0.4
0.3
0.2
0.1
0.0
−5.0
−2.5
0.0
2.5
5.0
x
colour
normal
t, df = 1
t, df = 3
t, df = 9
t, df = 27


Distribución t y normal estándar
probabilidad
1.00
0.75
0.50
0.25
0.00
−5.0
−2.5
0.0
2.5
5.0
x
colour
normal
t, df = 1
t, df = 3
t, df = 9
t, df = 27

Teorema

Sea Z una variable aleatoria normal estándar y V una variable aleatoria chi cuadrada con ν grados de libertad. Si Z y V son independientes, entonces la distribución de la variable aleatoria T, donde

$$T = \frac{Z}{\sqrt{V/\nu}}$$

está dada por la función de densidad

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

Ésta distribución se conoce como **distribución t** con ν grados de libertad.

Corolario

Sean $X_1, X_2, \cdots, X_n$ variables aleatorias independientes norma les con media μ y desviación estándar σ. Sea

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{y} \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Entonces la variable aleatoria $\frac{\bar{x}-\mu}{s/\sqrt{n}}$ tiene una distribución t con $\nu = n-1$ grados de libertad.

Ejemplo

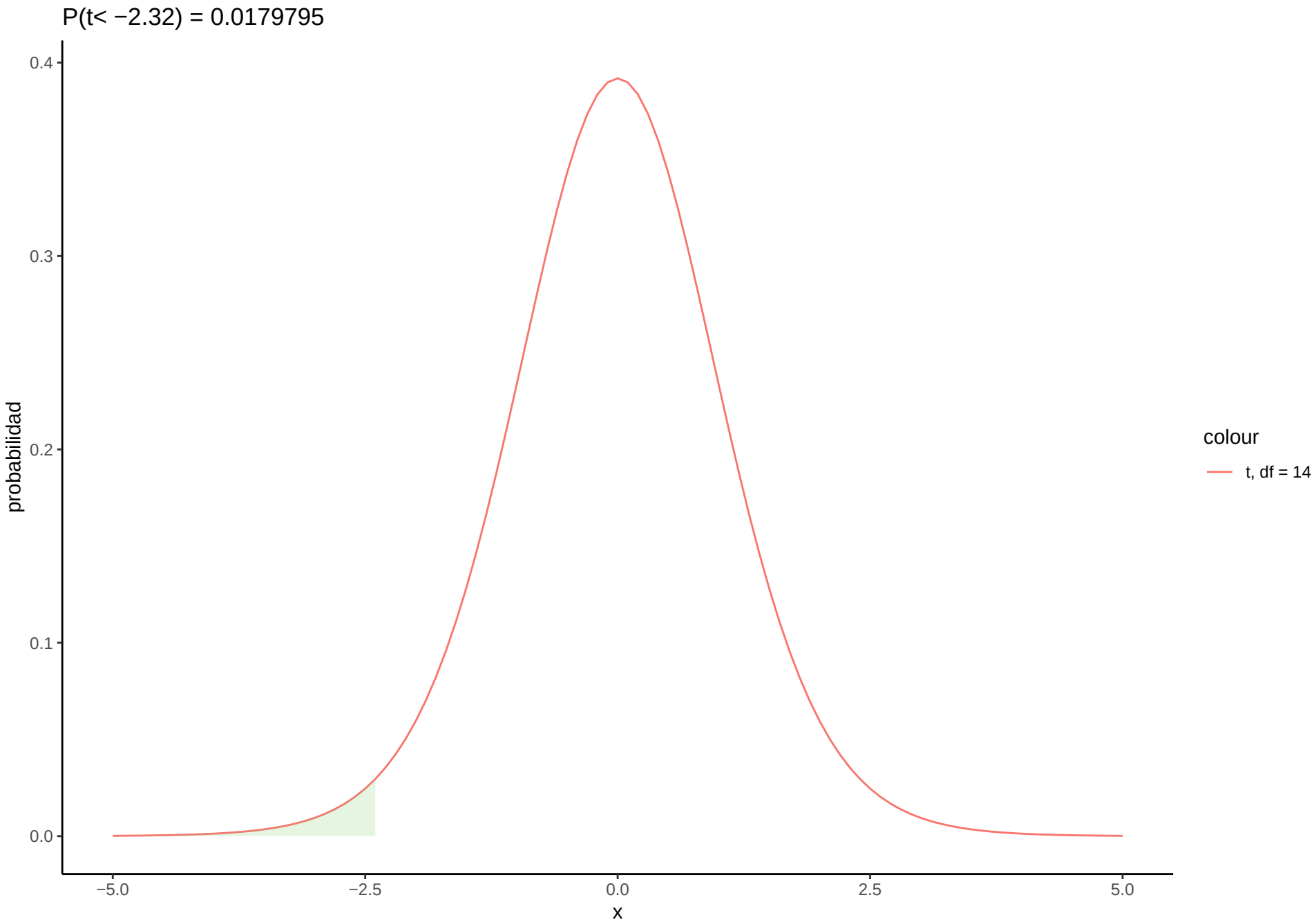
Suponga que usted tiene una técnica que puede modificar la edad a la cual los niños empiezan a hablar. En su localidad, el promedio de edad, en la cual un niño emite su primera palabra, es 13 meses. No conoce la desviación estandar poblacional. Usted aplica dicha técnica a una muestra de 15 niños. Los resultados son los siguientes:$8, 9, 10, 15, 18, 17, 12, 11, 7, 8, 10, 11, 8, 9, 12$. $n = 15$, $\bar{x} = 11.0$, desviación estándar $s = 3.34$. Si la media poblacional (verdadera) es 13 meses, ¿cuál es la probabilidad de encontrar un valor igual o menor de $\bar{x}$ de 11 meses?

Solución

Se tiene que $\mu = 13$, $s = 3.34$, $\bar{x} = 11$ y $n = 15$. Sustituyendo se obtiene

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{11 - 13}{3.34/\sqrt{15}} = -2.32$$

Por lo tanto $P(\bar{x} \leq 11) = P(t \leq -2.32) = 0.0179795$.


P(t< −2.32) = 0.0179795
0.4
0.3
0.2
0.1
0.0
probabilidad
−5.0
−2.5
0.0
2.5
5.0
x
colour
t, df = 14

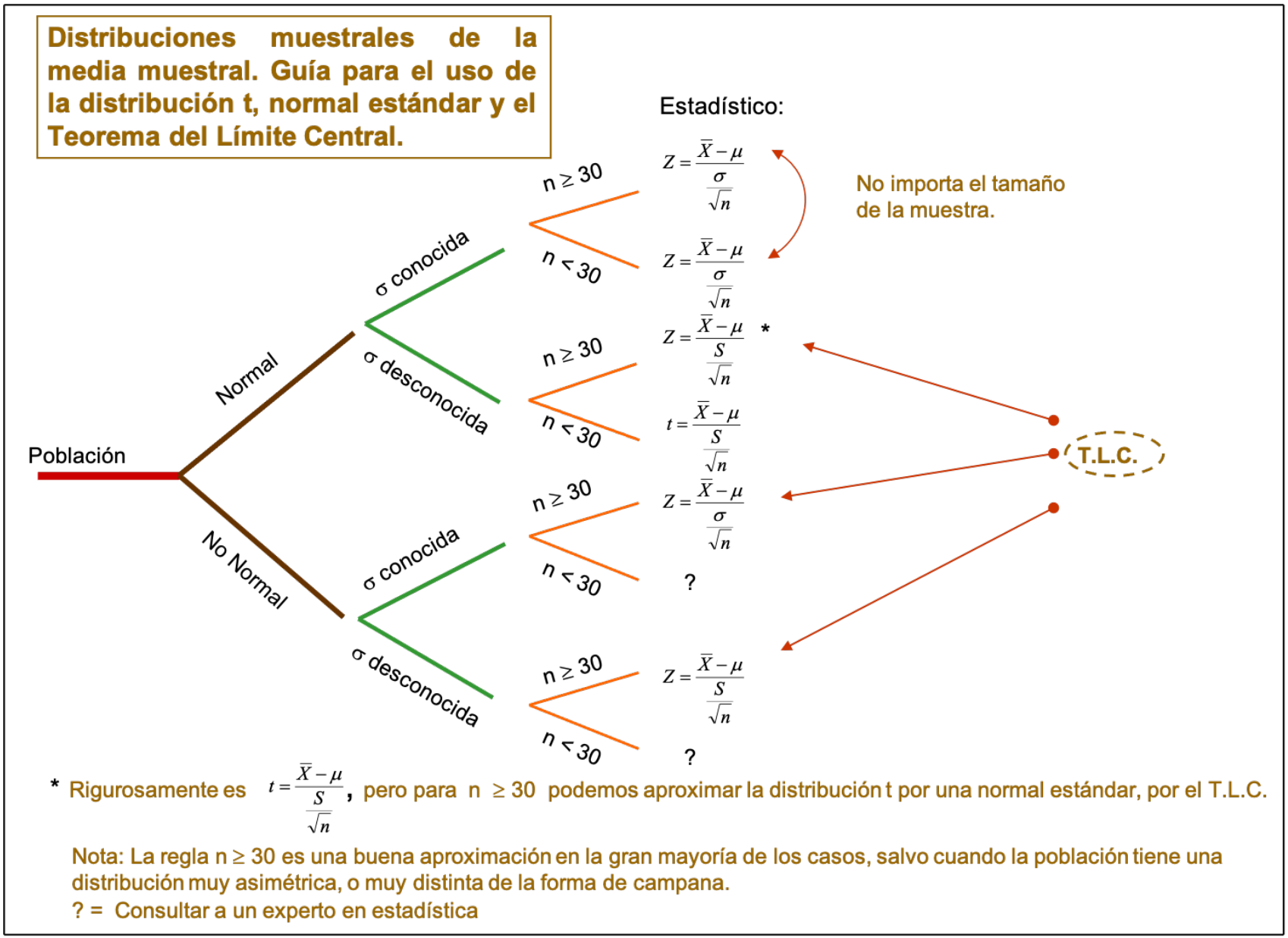


Figure 3: Distribuciones muestrales de la media

Distribución normal y las poblaciones discretas.

Aplicaciones

La distribución de normal se emplea muchas veces como una aproximación de valores en una población discreta. En situaciones, debe tenrse especial cuidado para asegurar que las probabilidades se calculan de manera precisa.

Considérese el siguiente ejemplo: Se sabe que el Coeficiente de Inteligencia (CI) de una población está distribuido normalmente en forma aproximada con $\mu = 100$ y $\sigma = 15$. ¿Cuál es la probabilidad de que un individuo seleccionado al azar tenga un CI de por lo menos 125? Si se hace $X = IC$ de una persona elegida al azar, deseamos $P(X \geq 125)$.

La tentación aquí es estandarizar como en los ejemplos anteriores. Sin embargo, la población del CI es discreta en realidad, ya que los CI son de valor entero, y la curva normal es una aproximación a un histograma de probabilidad discreta.

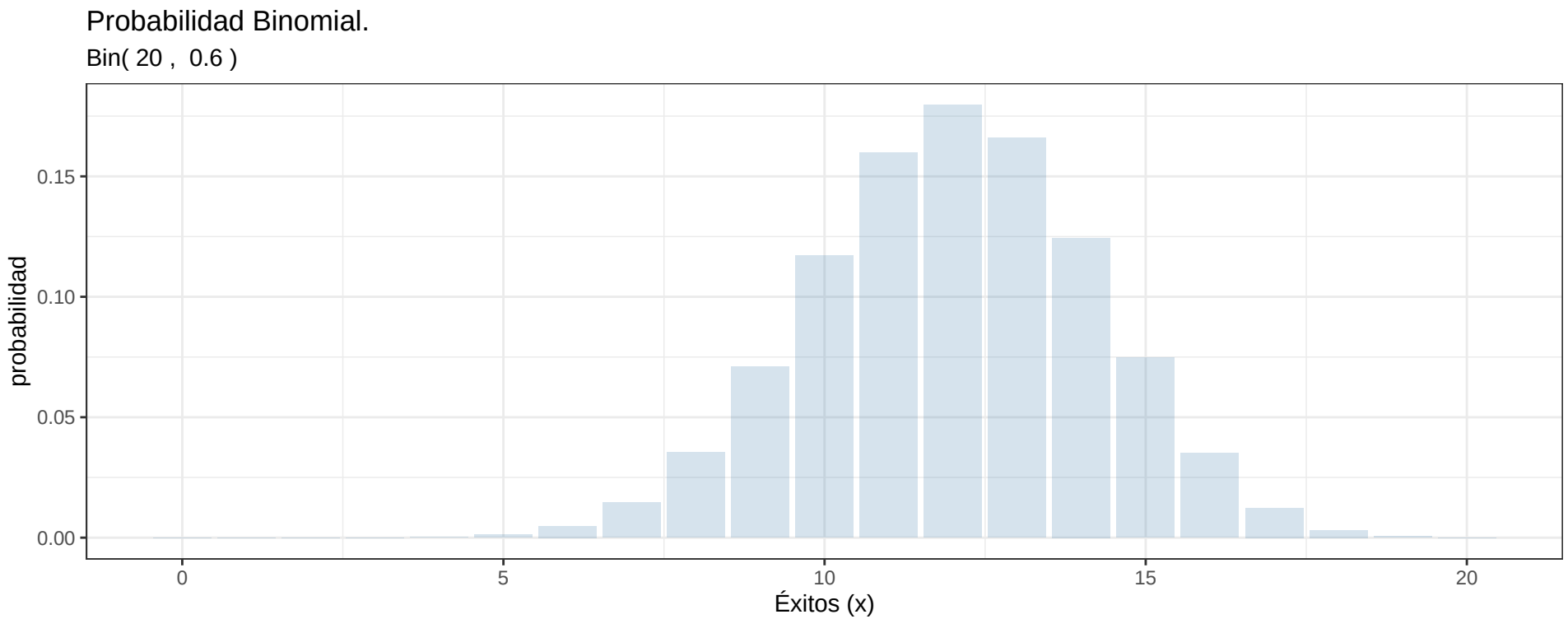
Los rectángulos del histograma están centrados como enteros, y los CI de por lo menos 125 corresponden a rectángulos que se inician en 124.5. En realidad deseamos $P(X \geq 124.5)$, que ahora se puede estandarizar para obtener $P(Z \geq 1.63) = 0.0516$.

Si hubieramos estandarizado $X \geq 125$, habríamos obtenido $P(Z \geq 1.67) = 0.0475$. La diferencia no es grande, pero la respuesta 0.0516 es más precisa. Análogamente, $P(X = 125)$ sería más apropiado por el área entre 124.5 y 125.5. Ya que el área bajo la curva normal arriba del valor único de 125 es cero.

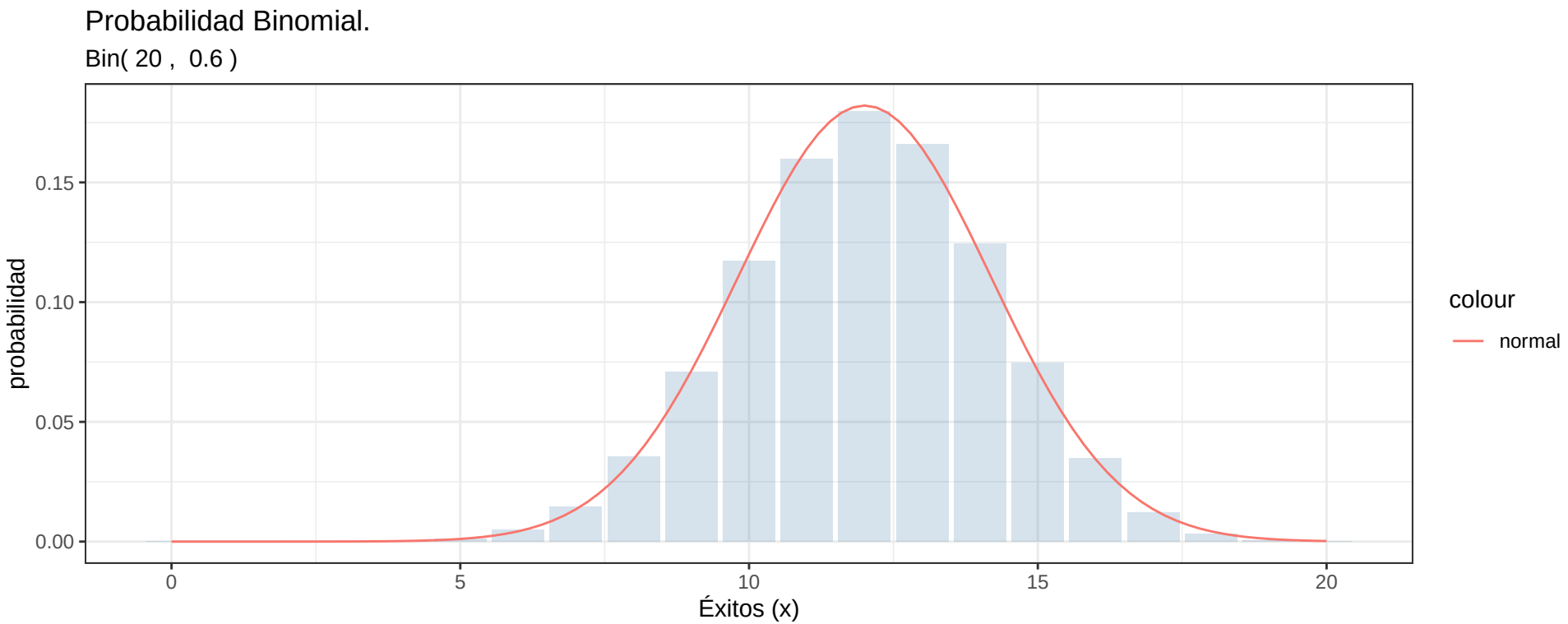
La corrección para la discretización de la distribución subyacente se llama con frecuencia **corrección de continuidad**. Es útil en la siguiente aplicación de la distribución normal

Aproximación de la distribución binomial a la distribución normal.

Recordemos que el valor medio y la desviación estándar de una variable aleatoria X binomial son $\mu_X = np$ $\sigma_X = \sqrt{npq}$, respectivamente. El siguiente histograma muestra una distribución binomial con $n = 20$, $p = 0.6$ (así que $\mu = 12$, $\sigma = [20(0.6)(0.4)]^{1/2} = 2.19$).



Una curva normal con valor medio y desviación estándar igual a los valores correspondientes para la distribución binomial se ha sobrepuesto en el histograma de probabilidad. Aun cuando el histograma esta un poco sesgado (porque $p \neq 0.5$), la curva normal da una buena aproximación, en especial en la parte media de la figura.



x	Binomial	Normal
0	0.000	0.000
1	0.000	0.000
2	0.000	0.000
3	0.000	0.000
4	0.000	0.000
5	0.001	0.001
6	0.005	0.005
7	0.015	0.014
8	0.035	0.035
9	0.071	0.072
10	0.117	0.120
11	0.160	0.163
12	0.180	0.181
13	0.166	0.163
14	0.124	0.120
15	0.075	0.072
16	0.035	0.035
17	0.012	0.014
18	0.003	0.005
19	0.000	0.001
20	0.000	0.000

Para el caso binomial: Tenemos que $X \sim Binom(n, p)$

$$P(X = x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}.$$

Si definimos $\mu = np$, $\sigma = \sqrt{np(1-p)}$, podemos aproximar a X mediante la variable $Y \sim N(\mu, \sigma)$ si calculamos de la siguiente forma:

$$P(X = x) \approx P(x - .5 \leq Y \leq x + .5)$$

$$= P\left(\frac{[x - .5] - \mu}{\sigma} \leq Z \leq \frac{[x + .5] - \mu}{\sigma}\right)$$

El área de cualquier rectángulo (probabilidad de cualquier valor de X particular), excepto los de las colas de los extremos, se puede aproximar con presición mediante el área de la curva normal correspondiente.

Por ejemplo, $P(X = 10) = b(X = 10; n = 20, p = 0.6) = 0.117$, mientras que el área bajo la curva normal entre 9.5 y 10.5 es $P(-1.14 < Z < -0.68) = 0.120$.

Más generalmente, mientras el histograma de probabilidad binomial no esté demasiado sesgado, las probabilidades binomiales se pueden aproximar bien por áreas de curva normal. Se dice entonces que X tiene aproximadamente una distribución normal.

Proposición : Sea X una V.A. Binomial basada en n intentos con probabilidad de éxito p. Entonces, si el histograma de probabilidad binomial no está demasiado sesgado, X tiene aproximadamente una distribución normal con $\mu = np$ y $\sigma = \sqrt{npq}$.

En particular, para $x =$ un valor posible de X, $P(X \leq x) = B(x; n, p) \approx$ (área bajo la curva normal estándar a la izquierda de $x + 0.5$), es decir

$$P(X \leq x) = B(x; n, p) \approx \Phi\left(\frac{x + 0.5 - np}{\sqrt{np(1-p)}}\right)$$

En la práctica la aproximación es adecuada si $np \geq 5$ y $n(1-p) \geq 5$.

Papel De Probabilidad Normal

La gráfica de papel de Probabilidad Normal, o simplemente gráfica de probabilidad normal, es un procedimiento útil para verificar si un conjunto de datos puede ser adecuadamente modelado por una distribución normal (Bondad de Ajuste). Este procedimiento consiste en construir una gráfica en el plano cartesiano, en donde, *en el eje horizontal se grafican los datos* y *en el eje vertical la probabilidad empírica (acumulada)* de los datos sobre una *escala de probabilidad normal*.

Es decir, es una gráfica que representa la distribución normal acumulada de los datos sobre una escala de probabilidad normal. Para construir la gráfica de probabilidad normal, deben disponerse los datos en orden ascendente y dibujar el $k-$*ésimo* de estos datos ordenados contra su punto de probabilidad acumulada $Pk = (k - 1/2)/n$ sobre papel de probabilidad normal. Si la distribución de los datos es normal, esta gráfica deberá parecer una línea recta.

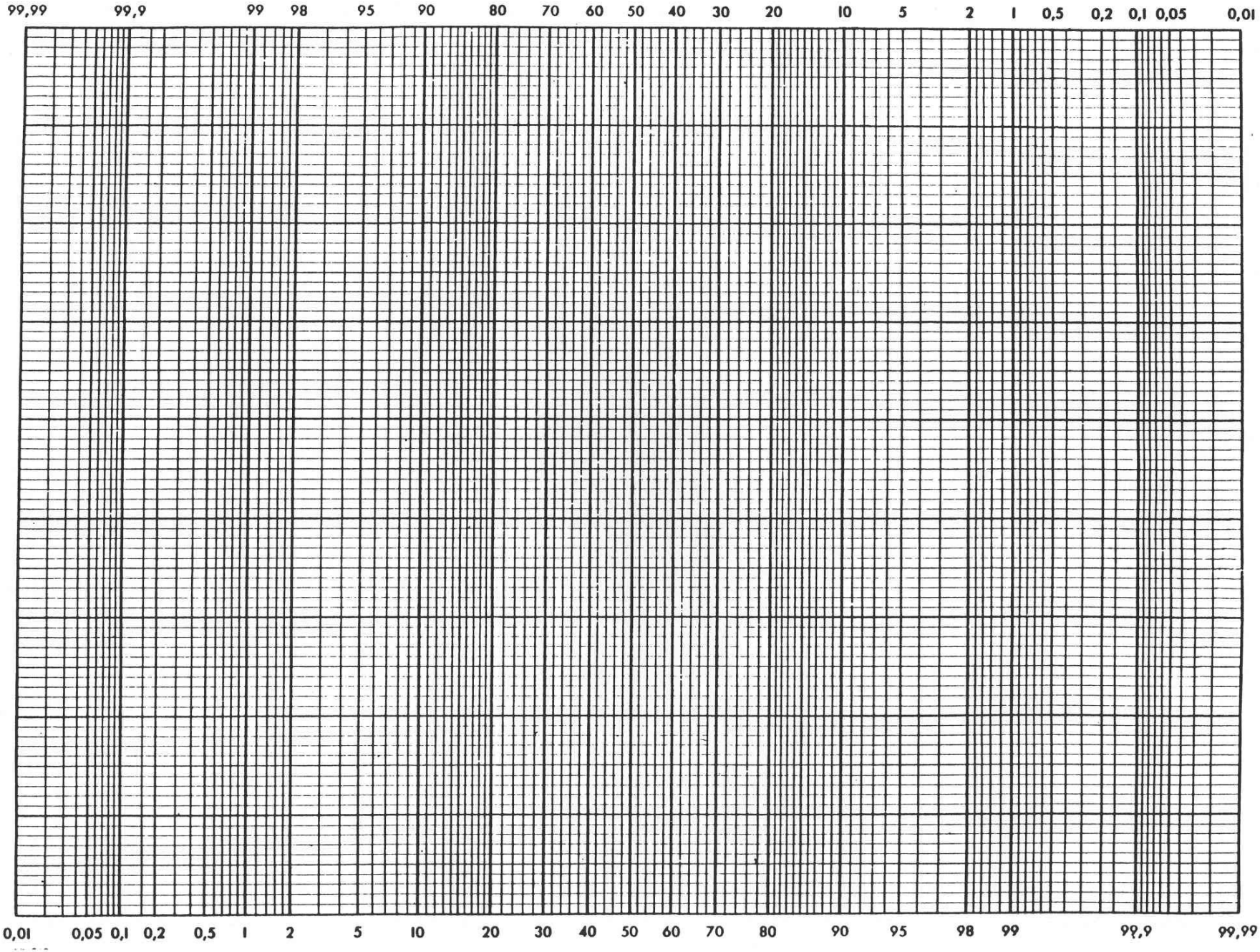
Para ejemplificar, considere los $n = 25$ datos siguientes.

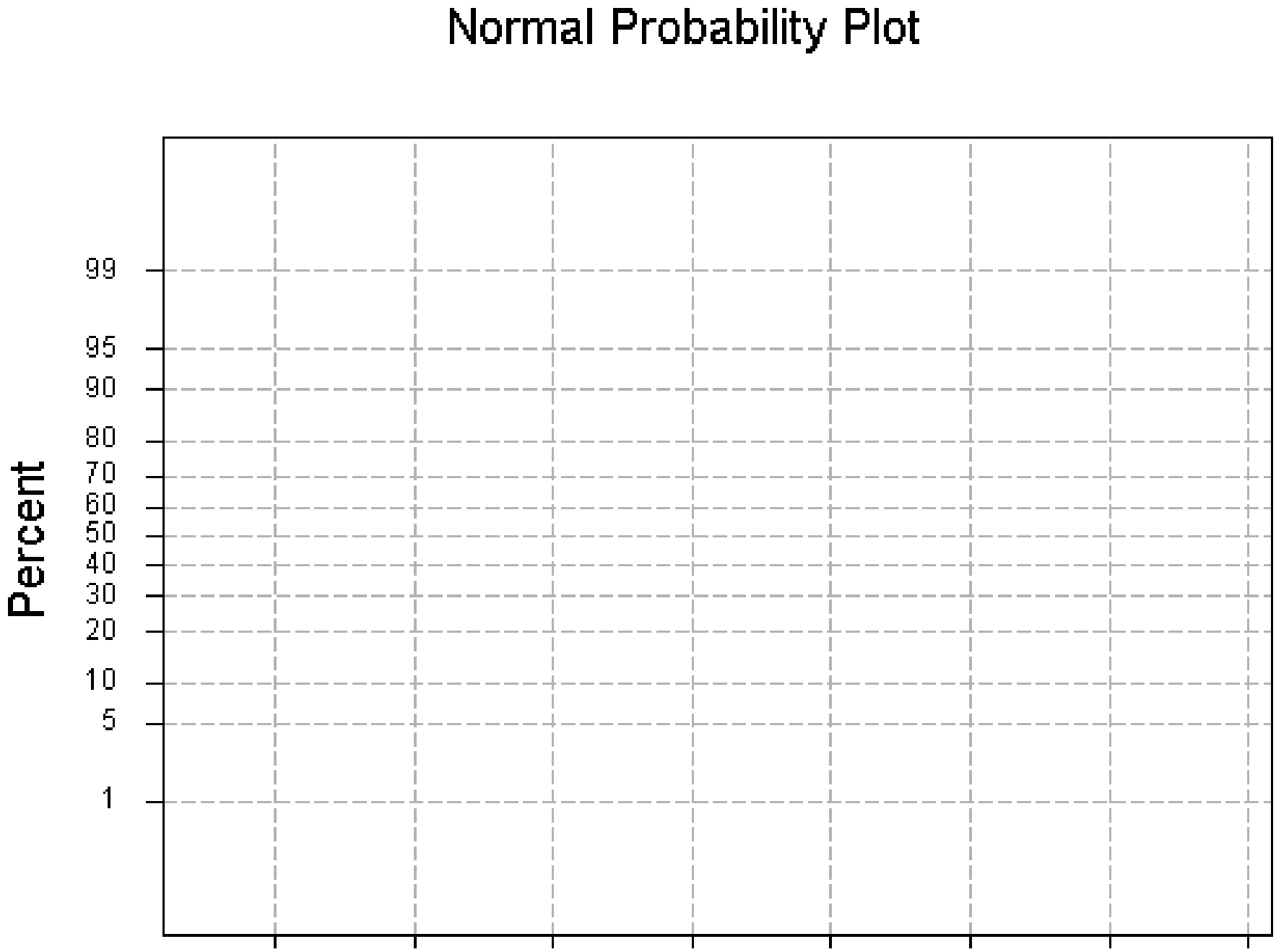
-0.8	3.4	-2.8	-2.6	1.2
2.6	0.4	0.2	5.2	1.4
-3.4	0.4	0.2	-0.8	1.6
-2.8	-3.8	-3.4	-2.6	2.6
1.4	1.4	0.4	4.2	-3.6

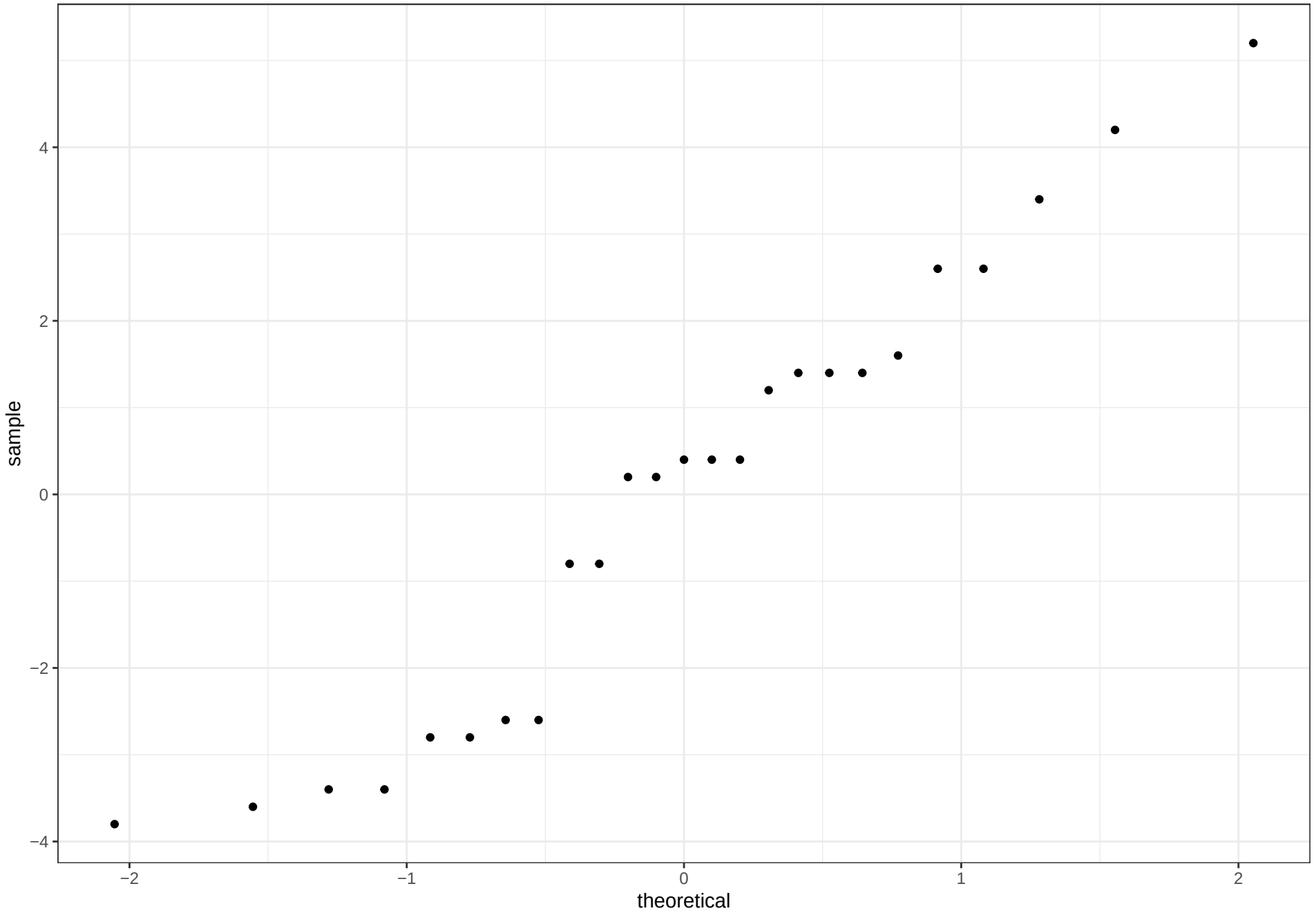
Ahora procederemos a ordenar de menor a mayor y calculamos el valor de P_k para el $k-$ésimo valor ordenado correspondiente con la expresión

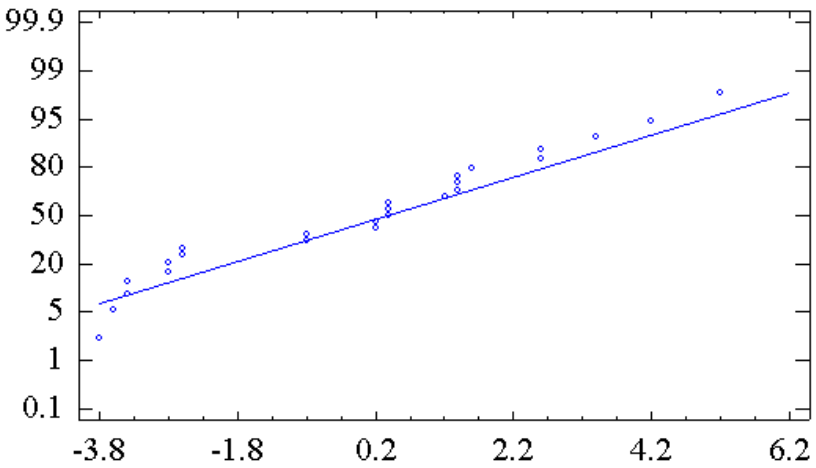
$$P_k = \frac{(k - 1/2)}{n}$$

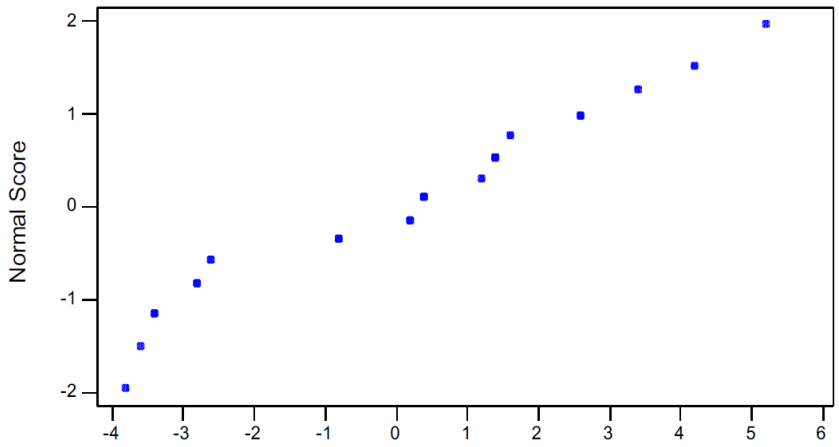
k	x	Pk	Porcentaje
1	-3.8	0.02	2
2	-3.6	0.06	6
3	-3.4	0.10	10
4	-3.4	0.14	14
5	-2.8	0.18	18
6	-2.8	0.22	22
7	-2.6	0.26	26
8	-2.6	0.30	30
9	-0.8	0.34	34
10	-0.8	0.38	38
11	0.2	0.42	42
12	0.2	0.46	46
13	0.4	0.50	50
14	0.4	0.54	54
15	0.4	0.58	58
16	1.2	0.62	62
17	1.4	0.66	66
18	1.4	0.70	70
19	1.4	0.74	74
20	1.6	0.78	78
21	2.6	0.82	82
22	2.6	0.86	86
23	3.4	0.90	90
24	4.2	0.94	94
25	5.2	0.98	98

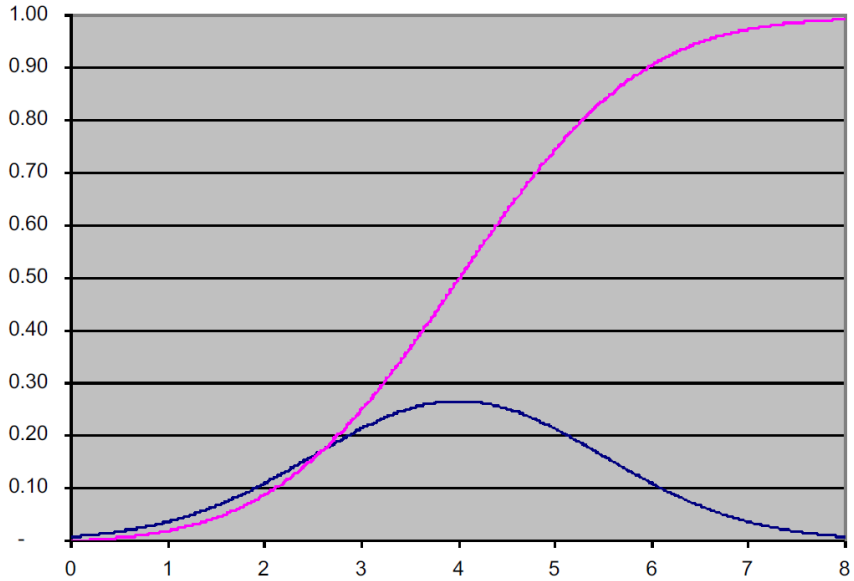

99,99 99,9 99 98 95 90 80 70 60 50 40 30 20 10 5 2 1 0,5 0,2 0,1 0,05 0,01
0,01 0,05 0,1 0,2 0,5 1 2 5 10 20 30 40 50 60 70 80 90 95 98 99 99,9 99,99

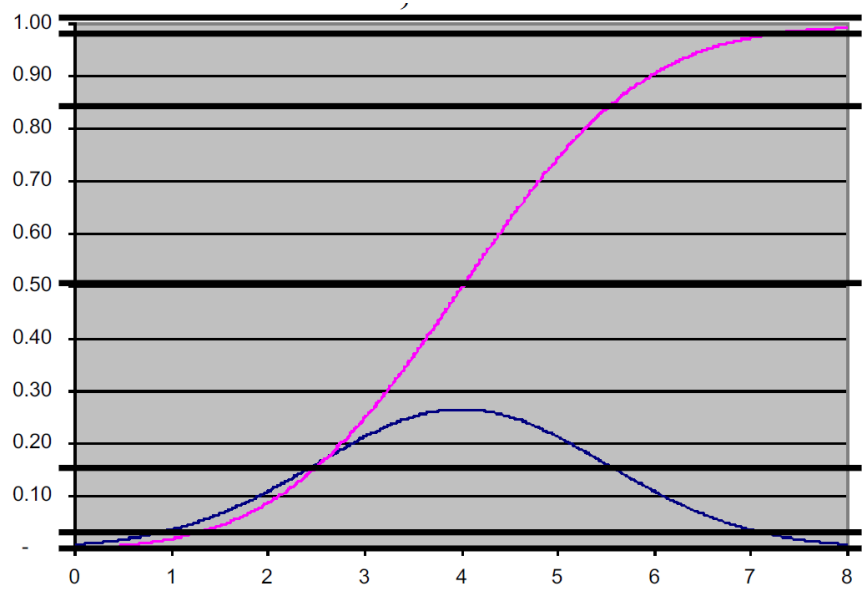

Normal Probability Plot
Percent
99
95
90
80
70
60
50
40
30
20
10
5
1

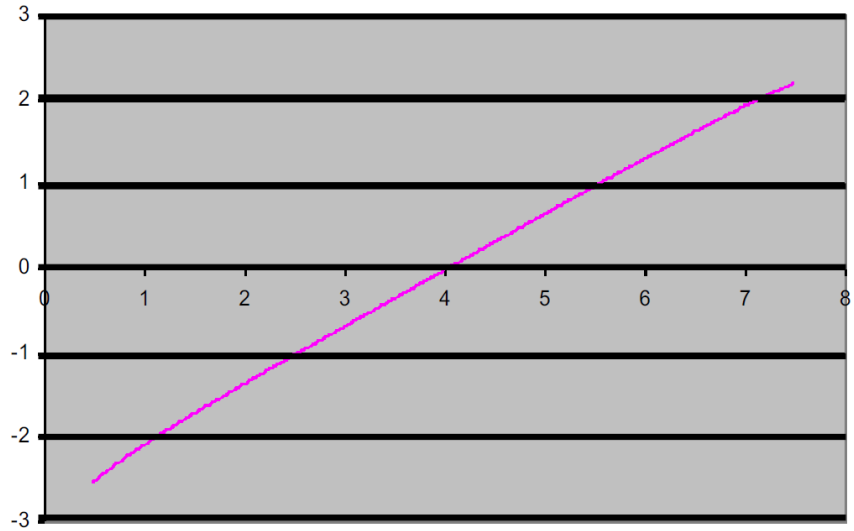

sample
theoretical
4
2
0
−2
−4
−2
−1
0
1
2


99.9
99
95
80
50
20
5
1
0.1
-3.8
-1.8
0.2
2.2
4.2
6.2


Normal Score
2
1
0
-1
-2
-4
-3
-2
-1
0
1
2
3
4
5
6





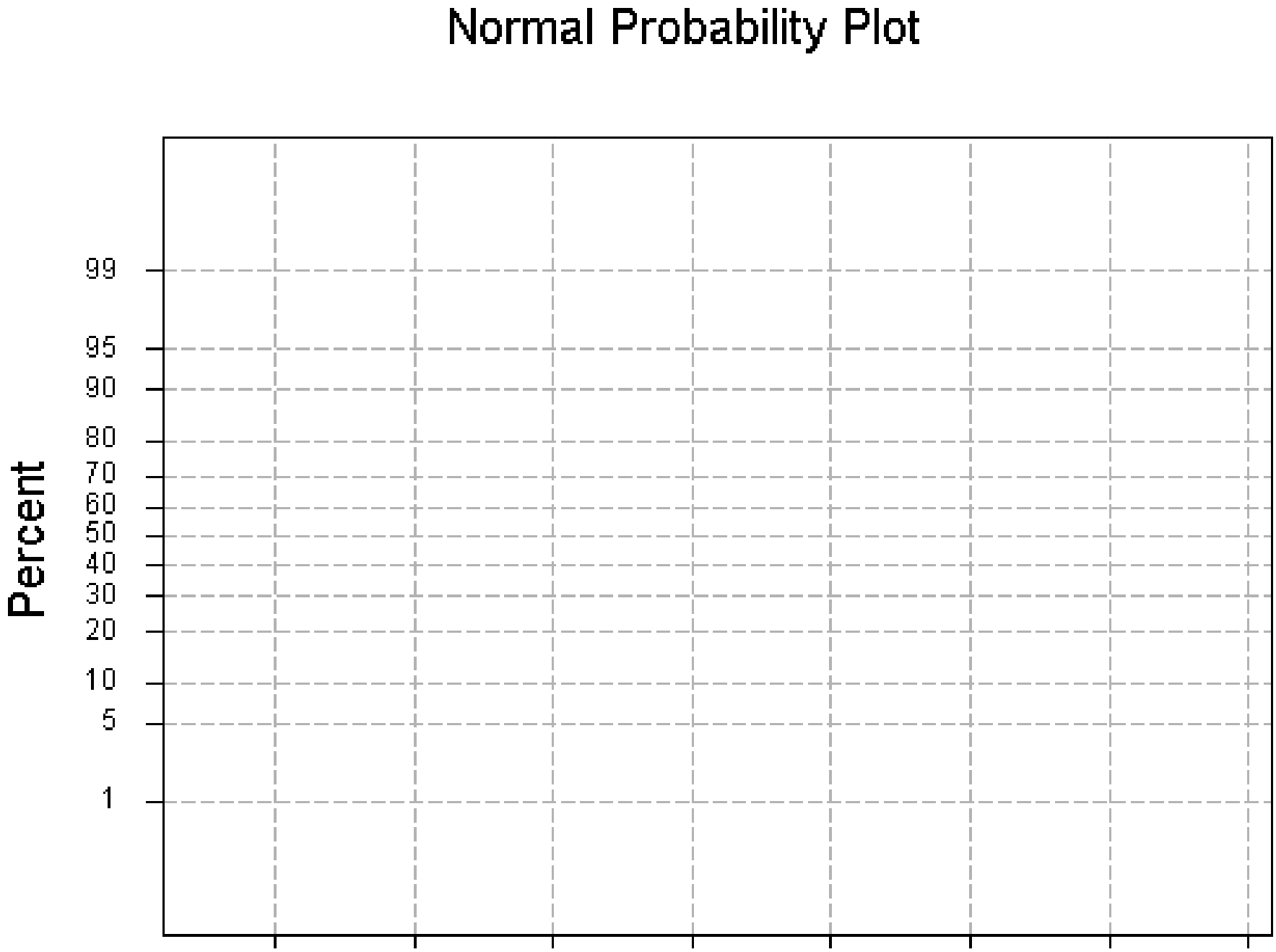


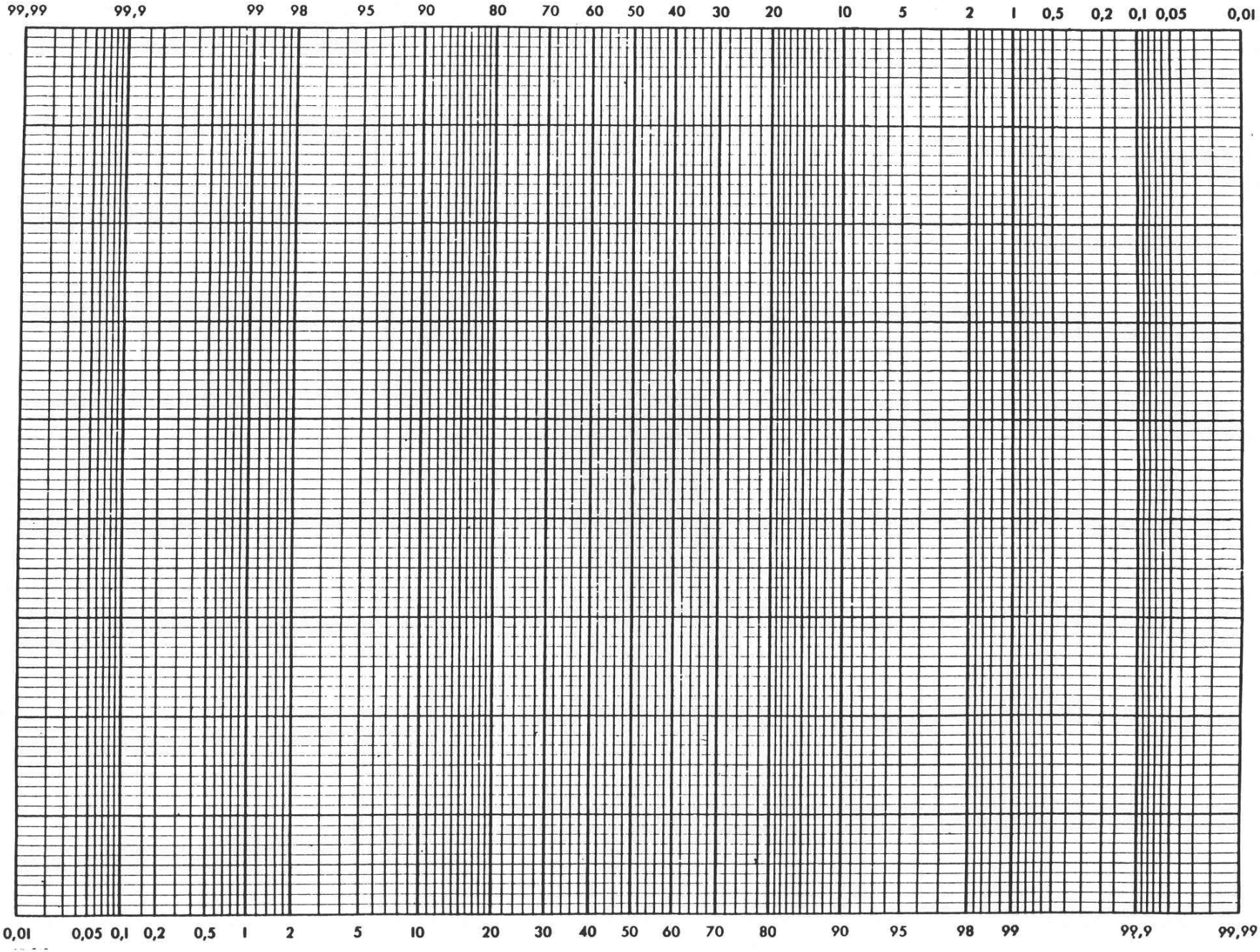
$$P\left[\mu-\sigma<X<\mu+\sigma\right]=0.683$$
$$P\left[\mu-2\sigma<X<\mu+2\sigma\right]=0.954$$
$$P\left[\mu-3\sigma<X<\mu+3\sigma\right]=0.997$$

Ejercicio de Papel de probabilidad Normal

Se prueba la duración de un componente electrónico bajo condiciones de temperatura alta para acelerar el mecanismo de falla. A continuación se proporciona el tiempo de falla (en horas) de 20 componentes seleccionados al azar. Haga una gráfica de los datos sobre papel de probabilidad normal. ¿El tiempo de falla parece tener una distribución normal?

176.1	24.7	34.9	133.8
76.6	55.0	122.8	99.6
150.4	73.0	90.6	131.5
197.6	124.5	2.4	40.4
35.3	155.7	46.0	40.4


Normal Probability Plot
Percent
99
95
90
80
70
60
50
40
30
20
10
5
1


99,99 99,9 99 98 95 90 80 70 60 50 40 30 20 10 5 2 1 0,5 0,2 0,1 0,05 0,01
0,01 0,05 0,1 0,2 0,5 1 2 5 10 20 30 40 50 60 70 80 90 95 98 99 99,9 99,99

Gráfica de Probabilidad normal en R

Práctica

- Realice la gráfica de probabilidad normal del ejemplo anterior.
- Simule con *semilla fija* datos con distribución Normal, Exponencial, Uniforme, t, Ji Cuadrada. Esto realícelo para tamaños de muestra de 15, 30, 50 y 100.
- ¿Qué conclusiones puede obtener a partir de esta práctica?

Ver la siguiente aplicación en Shiny qqnorm

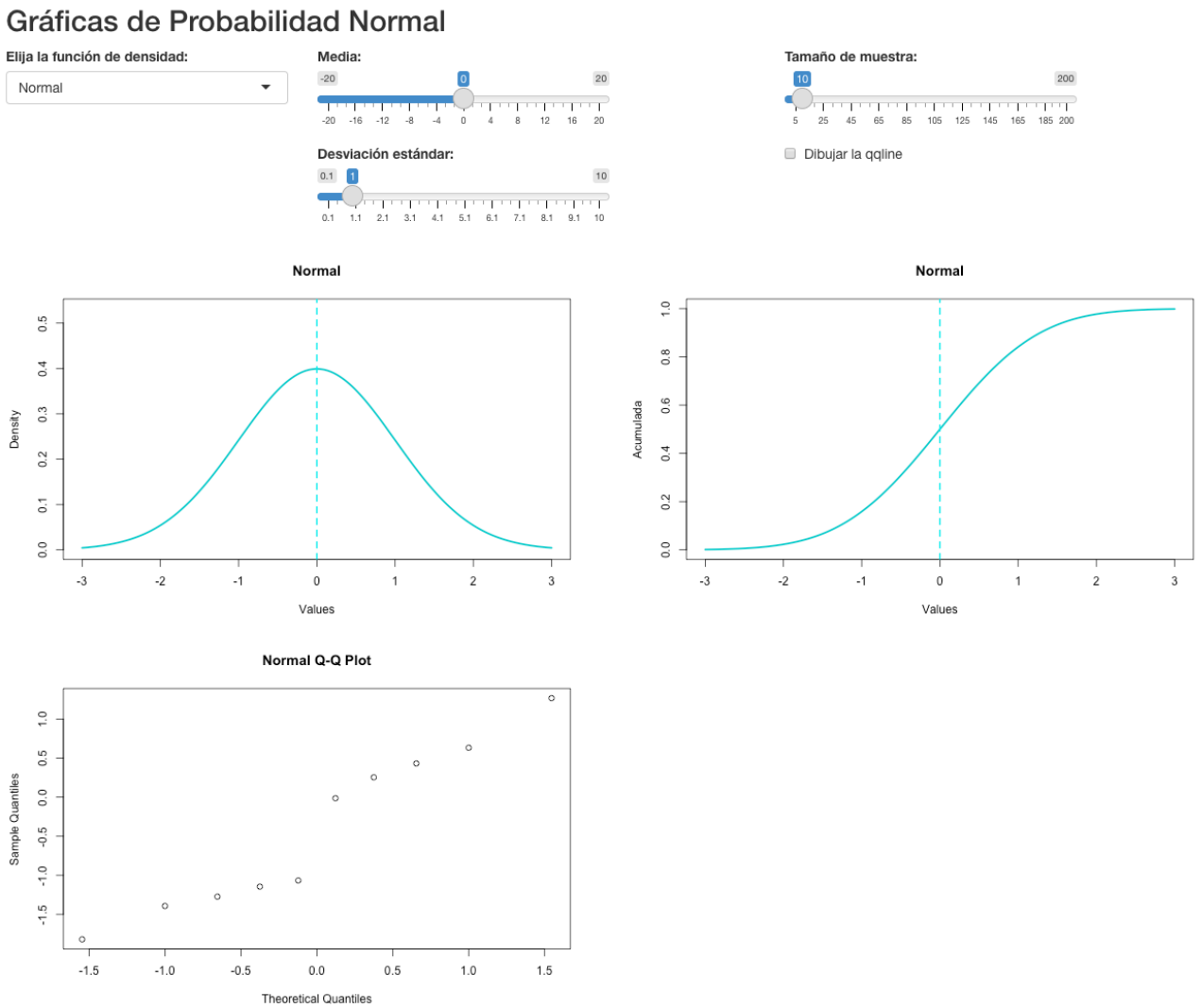


Figure 4: Pantalla de Shiny app QQnorm

¿Cómo se ven otras distribuciones muestrales?

Ver el archivo **distr_muestral_pequena2.R** en
https://rstudio.cloud/content/4731243
Simule la distribución muestral de la suma, mínimo y varianza para una población dada. La extracción será con reemplazo y con tamaño de muestra n fijo. Grafique la distribución muestral.

Distribución Ji-cuadrada (Chi-Square)

Hemos visto que el conocer la varianza σ^2 resulta fundamental para procedimientos de distribución muestral de la media, así como para procedimientos de inferencia estadística.

Existen muchas aplicaciones prácticas en donde σ^2 es el objetivo primario de la investigación experimental. (Precisión en el llenado de bolsas). En estos casos σ^2 adquiere una mayor importancia que la media de la población.

Las partes producidas por un proceso de manufactura deben ser producidas con un mínimo de variabilidad para reducir el número de productos fuera del rango aceptable (defectuosos). En general se desea mantener una varianza mínima en las características de calidad de un producto industrial para alcanzar el control del proceso y minimizar el porcentaje de productos de baja calidad.

La varianza muestral

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n-1}.$$

Es un estimador insesgado de la varianza de la población σ^2. La distribución muestral de s^2, generada mediante muestras repetidas, es una distribución de probabilidad que empieza en $s^2 = 0$ (ya que no puede ser negativa) con media igual a σ^2. La distribución **no es simétrica**.

La forma de la distribución depende del número de datos, así como de la forma de la distribución de origen.

En el caso de población normal

Si la población de origen de las muestras es normal, entonces la distribución estandarizada que se obtiene es la Ji- cuadrada, calculada como en la siguiente expresión:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

Relación de la Ji-cuadrada con la distribución normal

Si $Z \sim N(0,1)$, entonces Z^2 tiene la distribución gama especial a la que nos referimos como la distribución $ji - cuadrada$ con $\nu = 1$ **grado de libertad**. La Ji-cuadrada es importante en problemas de muestreo de poblaciones normales. En general si $Z_i \sim N(0,1)$, $i = 1, 2, \ldots, n$ independientes, entonces

$$X = Z_1^2 + Z_2^2 + \cdots + Z_n^2$$

entonces $X \sim \chi^2_n$.

Distribución Ji-cuadrada

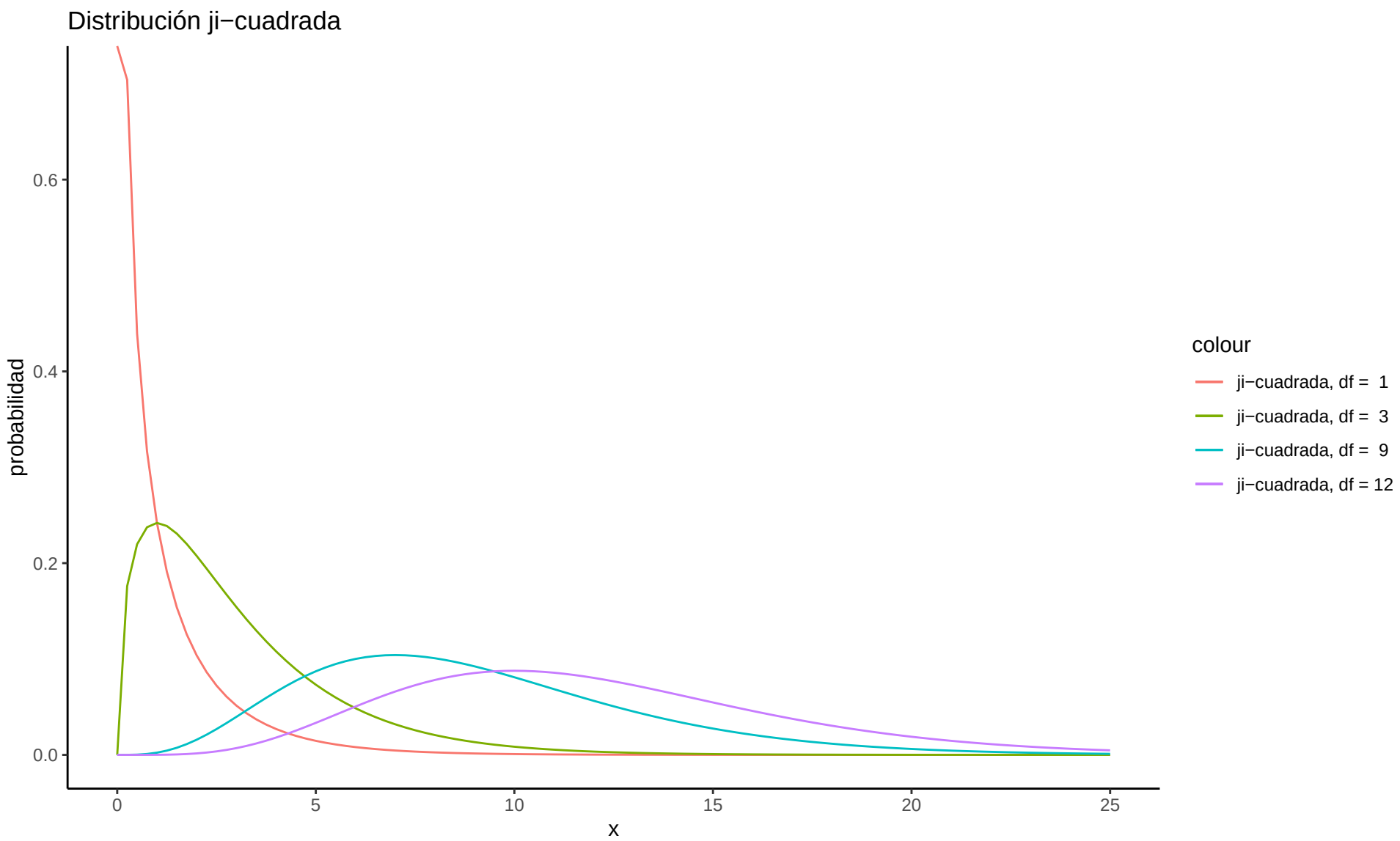
Una variable aleatoria x tiene una distribución ji-cuadrada (χ^2) con ν (nu) grados de libertad, si su densidad está dada por:

$$f(x) = \frac{1}{2^{\frac{\nu}{2}}\Gamma(\nu/2)} x^{\frac{\nu-2}{2}} e^{-x/2} \text{ para } x > 0$$

Función Gama

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} dy \qquad \text{para } \alpha > 0$$

Casos importantes: $\Gamma(1/2) = \sqrt{\pi}$
$\Gamma(k) = (k-1)!$ para k entero.


Distribución ji−cuadrada
probabilidad
0.0
0.2
0.4
0.6
0
5
10
15
20
25
x
colour
ji−cuadrada, df = 1
ji−cuadrada, df = 3
ji−cuadrada, df = 9
ji−cuadrada, df = 12

Estimación de Varianzas

Teorema. Si s^2 es la varianza de una muestra aleatoria de tamaño n tomada de una población normal cuya varianza es σ^2, entonces:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

es el valor de una variable aleatoria que tiene distribución Ji-cuadrada con parámetro $\gamma = \nu - 1$ grados de libertad.

Exactamente 95% de una distribución chi cuadrada cae entre $\chi^2_{0.975}$ y $\chi^2_{0.025}$. Un valor χ^2 que cae a la derecha de $\chi^2_{0.975}$ no tiene probabilidades de ocurrir, a menos que el valor de σ^2 que supusimos sea demasiado pequeño. Lo mismo sucede con un valor χ^2 que cae a la izquierda de $\chi^2_{0.025}$, el cual tampoco es probable que ocurra, a menos que el valor de σ^2 que supusimos sea demasiado grande. En otras palabras, es posible tener un valor χ^2 a la izquierda de $\chi^2_{0.025}$ o a la derecha de $\chi^2_{0.975}$ cuando el valor de σ^2 es correcto; pero si esto sucediera, lo más probable es que el valor de σ^2 que se supuso sea un error.

Ejemplo

Un fabricante de baterías para automóvil garantiza que su producto durará, en promedio, 3 años con una desviación estándar de 1 año. Si cinco de estas baterías tienen duraciones de $1.9, 2.4, 3.0, 3.5$ y 4.2 años, ¿el fabricante continuará convencido de que sus baterías tienen una desviación estándar de 1 año? Suponga que las duraciones de las baterías siguen una distribución normal.

Solución

Primero calculemos la media y varianza, muestrales:

$$\bar{x} = 3 \qquad \text{y} \qquad s^2 = \frac{(1.9-3)^2 + (2.4-3)^2 + \cdots + (4.2-3)^2}{4} = 0.815$$

Entonces

$$\chi^2 = \frac{(4)(0.815)}{1} = 3.26$$

es un valor de una distribución chi cuadrada con 4 grados de libertad. Como 95% de los valores χ^2 con 4 grados de libertad caen entre 0.484 y 11.143, el valor calculado con $\sigma^2 = 1$ es razonable y, por lo tanto, el fabricante no tiene razones para sospechar que la desviación estándar no sea igual a 1 año.

Distribución F

Fisher-Snedecor

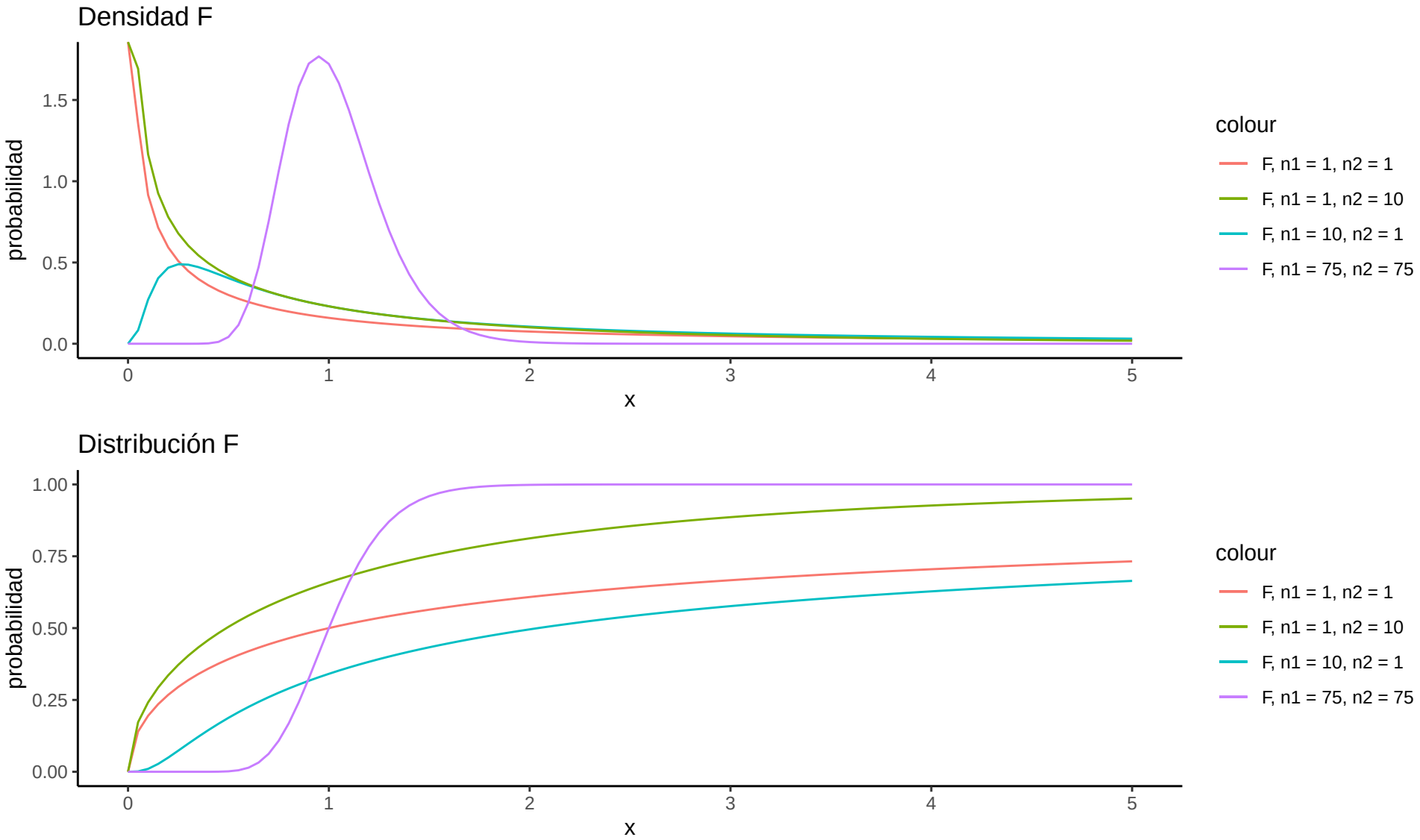
Sí $X_1 \sim \chi^2_{n_1}$ y $X_2 \sim \chi^2_{n_2}$ y son independientes y definimos

$$X = \frac{X_1/n_1}{X_2/n_2}$$

entonces $X \sim F_{n_1,n_2}$ a veces denotada como $F(n_1, n_2)$ y su función de densidad está dada por

$$f(x) = \frac{\Gamma((n_1+n_2)/2)(n_1/n_2)^{n_1/2}x^{n_1/2-1}}{\Gamma(n_1/2)\Gamma(n_2/2)[(n_1/n_2)x+1]^{(n_1+n_2)/2}} \qquad x > 0,$$

para $n_1 = 1, 2, \ldots$ y $n_2 = 1, 2, \ldots$. La distribución F se usa para inferencia estadística de razones de varianzas de dos poblaciones normales. También se usa para inferencia estadística de razones de tasas de dos poblaciones exponenciales.


Densidad F
probabilidad
1.5
1.0
0.5
0.0
0
1
2
3
4
5
x
colour
F, n1 = 1, n2 = 1
F, n1 = 1, n2 = 10
F, n1 = 10, n2 = 1
F, n1 = 75, n2 = 75
Distribución F
probabilidad
1.00
0.75
0.50
0.25
0.00
0
1
2
3
4
5
x
colour
F, n1 = 1, n2 = 1
F, n1 = 1, n2 = 10
F, n1 = 10, n2 = 1
F, n1 = 75, n2 = 75