

# **Muestreo y Estadística**

## **Poblaciones y muestras, Estadístico, Teorema del límite central, test t y contraste de hipótesis**

Sebastián Ruiz-Blais, I-2024

# Contenidos

- Poblaciones y muestras
- Muestreo
- Estadístico
- Teorema del límite central
- Test t y contraste de hipótesis

# Poblaciones y muestras

- Una **población** es la totalidad de las observaciones que nos interesan
  - Por ejemplo, todos los estudiantes universitarios de Costa Rica o todas las especies de insectos en el planeta
- Una **muestra** es un subconjunto de la población
  - Por ejemplo, tomar un grupo de cincuenta estudiantes o un grupo de 500 insectos de distintos países

# Pregunta

- Algunos datos con respecto a la **población de insectos en el planeta**
  - Se estima que la población de insectos a nivel mundial ha decaído en un 65%
  - Se han observado caídas de 76% para Alemania y entre 75 y 98% en Puerto Rico
  - Se estima que la cantidad de especies que ha decaído es de 40%
- **¿Cómo cree usted que este tipo de estudios han sido realizados?**
  - ¿Cuál es la población y cuál es la muestra?
  - ¿Cómo se relacionan?

# Muestreo

- **Muestreo** es el proceso de obtener una muestra a partir de una población.
- Para evitar sesgos es importante tomar muestras de manera aleatoria



# Muestreo

- **Muestreo** es el proceso de obtener una muestra a partir de una población.
- Para evitar sesgos es importante tomar muestras de manera aleatoria
- **¿Cuáles serían posibles sesgos en una muestra?**



# Definición de muestra aleatoria

**Definición 8.3:** Sean  $X_1, X_2, \dots, X_n$  variables aleatorias independientes  $n$ , cada una con la misma distribución de probabilidad  $f(x)$ . Definimos  $X_1, X_2, \dots, X_n$  como una **muestra aleatoria** de tamaño  $n$  de la población  $f(x)$  y escribimos su distribución de probabilidad conjunta como

$$f(x_1, x_2, \dots, x_n) = f(x_1) f(x_2) \cdots f(x_n).$$

- Por ejemplo, cada  $X_i$  puede corresponder a personas en una encuesta o muestras de sangre en un estudio

# Estadístico

**Definición 8.4:** Cualquier función de las variables aleatorias que forman una muestra aleatoria se llama estadístico.

- **Ejemplo.** Si quisiéramos determinar la proporción de personas que apoya a un candidato presidencial, podríamos preguntar a toda la población y obtener su valor  $p$ . Sin embargo, como esto no es plausible, usualmente tomamos una muestra grande y obtenemos la probabilidad  $\hat{p}$  correspondiente a esa muestra.

# Medidas de localización

- Algunos estadísticos que se pueden obtener son las medidas de localización, que se utilizan para medir el centro del conjunto de datos

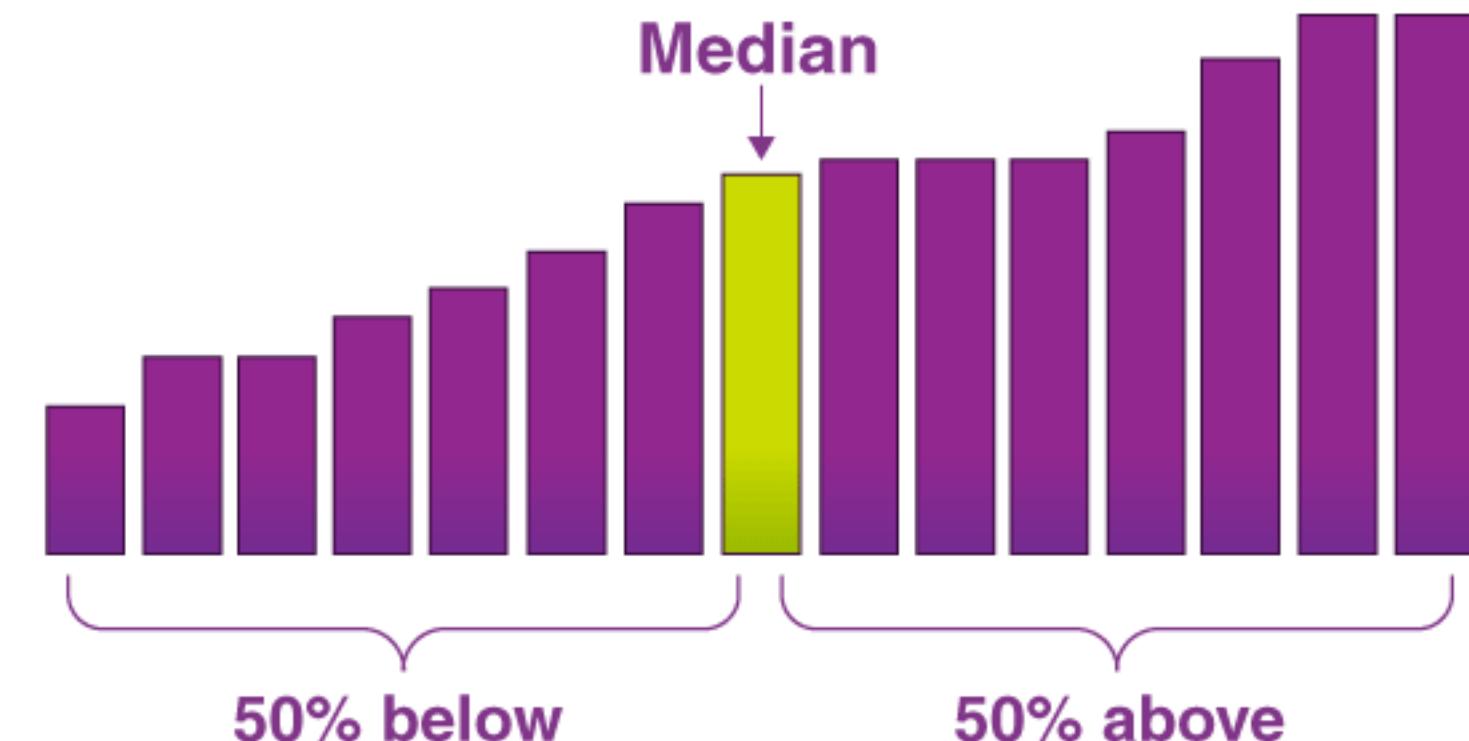
# Media, mediana y moda

Media muestral:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i.$$

Mediana muestral:

$$\tilde{x} = \begin{cases} x_{(n+1)/2}, & \text{si } n \text{ es impar} \\ \frac{1}{2}(x_{n/2} + x_{n/2+1}), & \text{si } n \text{ es par.} \end{cases}$$



La moda muestral es el valor que ocurre con mayor frecuencia en la muestra.

# Ejercicio

- Suponga que un conjunto de datos consta de las siguientes observaciones, en orden:
  - **0.32, 0.53, 0.28, 0.37, 0.47, 0.43, 0.36, 0.42, 0.38, 0.43**
- Obtenga la media, mediana y moda a partir de estos datos

# Varianza muestral

- Así como se tienen medidas de tendencia central, para comprender cómo se distribuyen los datos se obtiene la varianza muestral

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

# Ejercicio

- Una comparación de los precios de café en 4 tiendas de abarrotes de San Marcos, seleccionadas al azar, mostró aumentos en comparación con el mes anterior de 12, 15, 17 y 20 centavos por bolsa de una libra. Calcule la varianza de esta muestra aleatoria de aumentos de precio.

# Solución

Si calculamos la media de la muestra, obtenemos

$$\bar{x} = \frac{12 + 15 + 17 + 20}{4} = 16 \text{ centavos.}$$

Por lo tanto,

$$\begin{aligned}s^2 &= \frac{1}{3} \sum_{i=1}^4 (x_i - 16)^2 = \frac{(12 - 16)^2 + (15 - 16)^2 + (17 - 16)^2 + (20 - 16)^2}{3} \\&= \frac{(-4)^2 + (-1)^2 + (1)^2 + (4)^2}{3} = \frac{34}{3}.\end{aligned}$$

# Distribución muestral de medias

- De una población normal con media  $\mu$  y varianza  $\sigma^2$  se toma una muestra de  $n$  observaciones.
- Cada observación de la muestra tiene la misma distribución que la población de la cual se obtiene.
- La media de la muestra 
$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$
- La cual tiene una distribución normal con **media**  $\mu_{\bar{x}} = \mu$  y **varianza**  $\sigma_{\bar{x}}^2 = \sigma^2/n$

# Teorema del Límite Central

**Teorema del límite central:** Si  $\bar{X}$  es la media de una muestra aleatoria de tamaño  $n$ , tomada de una población con media  $\mu$  y varianza finita  $\sigma^2$ , entonces la forma límite de la distribución de

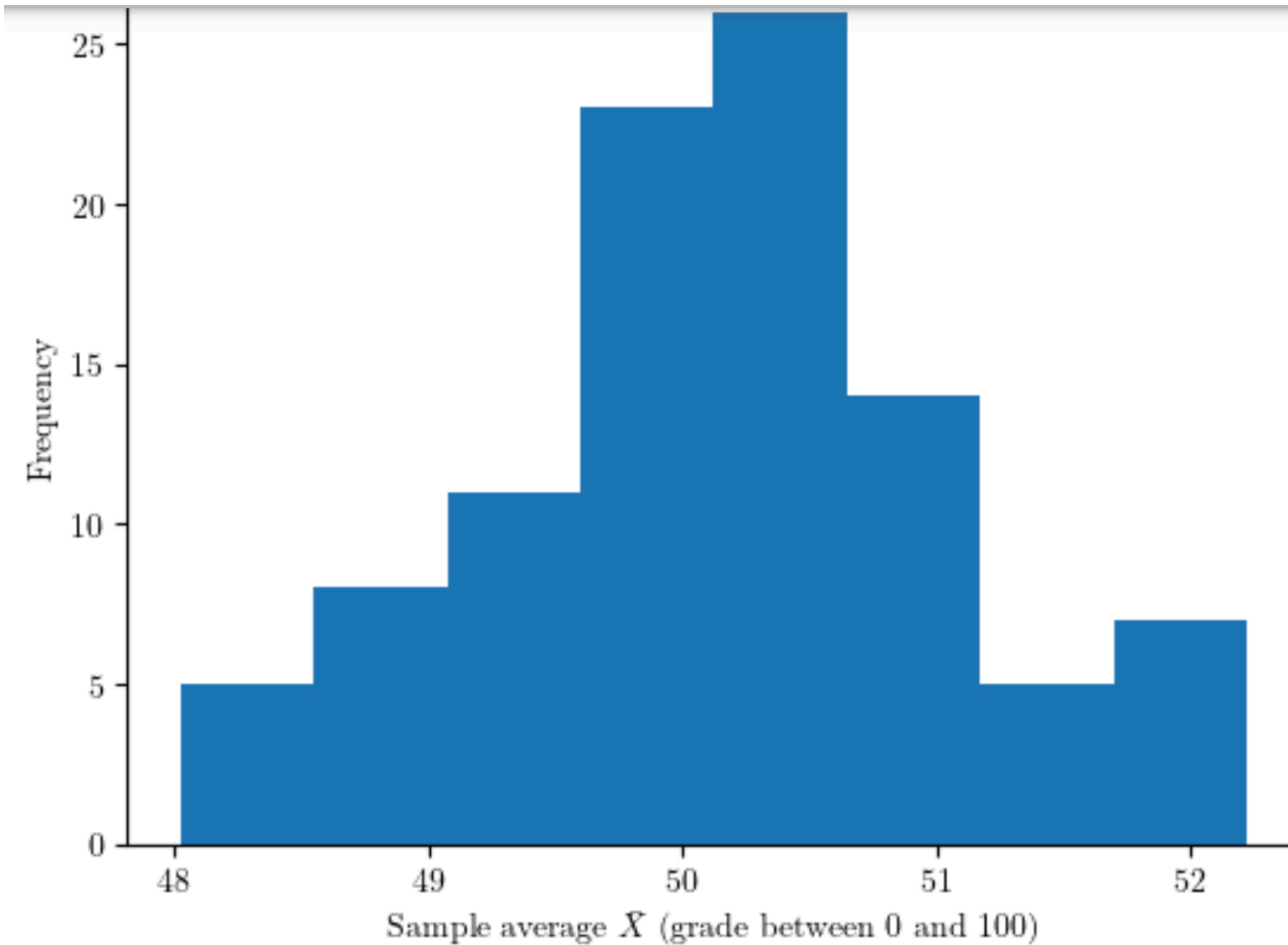
$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}},$$

a medida que  $n \rightarrow \infty$ , es la distribución normal estándar  $n(z; 0, 1)$ .

# Teorema del Límite Central

- Esto se cumple sin importar la distribución de la población.
- Usualmente la aproximación se cumple cuando  $n \geq 30$ , siempre que la distribución no sea muy asimétrica y para  $n < 30$  cuando la distribución de la población es similar a la distribución normal.
- Esta suposición de normalidad se hace más precisa conforme  $n$  es mayor.

# Experimento en Python



# Problema

- Una empresa de material eléctrico fabrica bombillas que tienen una duración que se distribuye aproximadamente en forma normal, con media de 800 horas y desviación estándar de 40 horas. Calcule la probabilidad de que una muestra aleatoria de 16 bombillas tenga una vida promedio de menos de 775 horas.

# Solución

- Por el teorema del límite central, la distribución muestral de  $\bar{X}$  será aproximadamente normal, con  $\mu_{\bar{X}} = 800$  y  $\sigma_{\bar{X}} = 40/\sqrt{16} = 10$ .
- En lo que corresponde a  $\bar{x} = 775$ , obtenemos que

$$z = \frac{775 - 800}{10} = -2.5$$

Por lo tanto:  $P(\bar{X} < 775) = P(Z < -2.5) = 0.0062$

# Distribución muestral de $\bar{X}_1 - \bar{X}_2$

Si se extraen al azar muestras independientes de tamaños  $n_1$  y  $n_2$  de dos poblaciones, discretas o continuas, con medias  $\mu_1$  y  $\mu_2$  y varianzas  $\sigma_1^2$  y  $\sigma_2^2$ , respectivamente, entonces la distribución muestral de las diferencias de las medias,  $\bar{X}_1 - \bar{X}_2$ , tiene una distribución aproximadamente normal, con media y varianza dadas por

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2 \text{ y } \sigma_{\bar{X}_1 - \bar{X}_2}^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

De aquí,

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}}$$

es aproximadamente una variable normal estándar.

# Problema

## Tiempo de secado de pinturas

- Se llevan a cabo dos experimentos independientes en los que se comparan dos tipos diferentes de pintura, el A y el B. Con la pintura tipo A se pintan 18 especímenes y se registra el tiempo (en horas) que cada uno tarda en secar. Lo mismo se hace con la pintura tipo B. Se sabe que la desviación estándar de la población de ambas es 1.0.
- Si se supone que los especímenes pintados se secan en el mismo tiempo medio con los dos tipos de pintura, calcule  $P(\bar{X}_A - \bar{X}_B > 1.0)$ , donde  $\bar{X}_A$  y  $\bar{X}_B$  son los tiempos promedio de secado para muestras de tamaño  $n_A = n_B = 18$

# Solución

- A partir de la distribución de muestreo de  $\bar{X}_A - \bar{X}_B$  sabemos que la distribución es aproximadamente normal con media:  $\mu_{\bar{X}_A - \bar{X}_B} = \mu_A - \mu_B = 0$

y varianza:  $\sigma_{\bar{X}_A - \bar{X}_B}^2 = \frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B} = \frac{1}{18} + \frac{1}{18} = \frac{1}{9}$

# Solución

- Ahora bien, la probabilidad que buscamos es  $P(\bar{X}_A - \bar{X}_B > 1.0)$ , que se aprecia en la parte sombreada de la figura
- Obtenemos el z:

$$z = \frac{1 - (\mu_A - \mu_B)}{\sqrt{1/9}} = \frac{1 - 0}{\sqrt{1/9}} = 3.0;$$

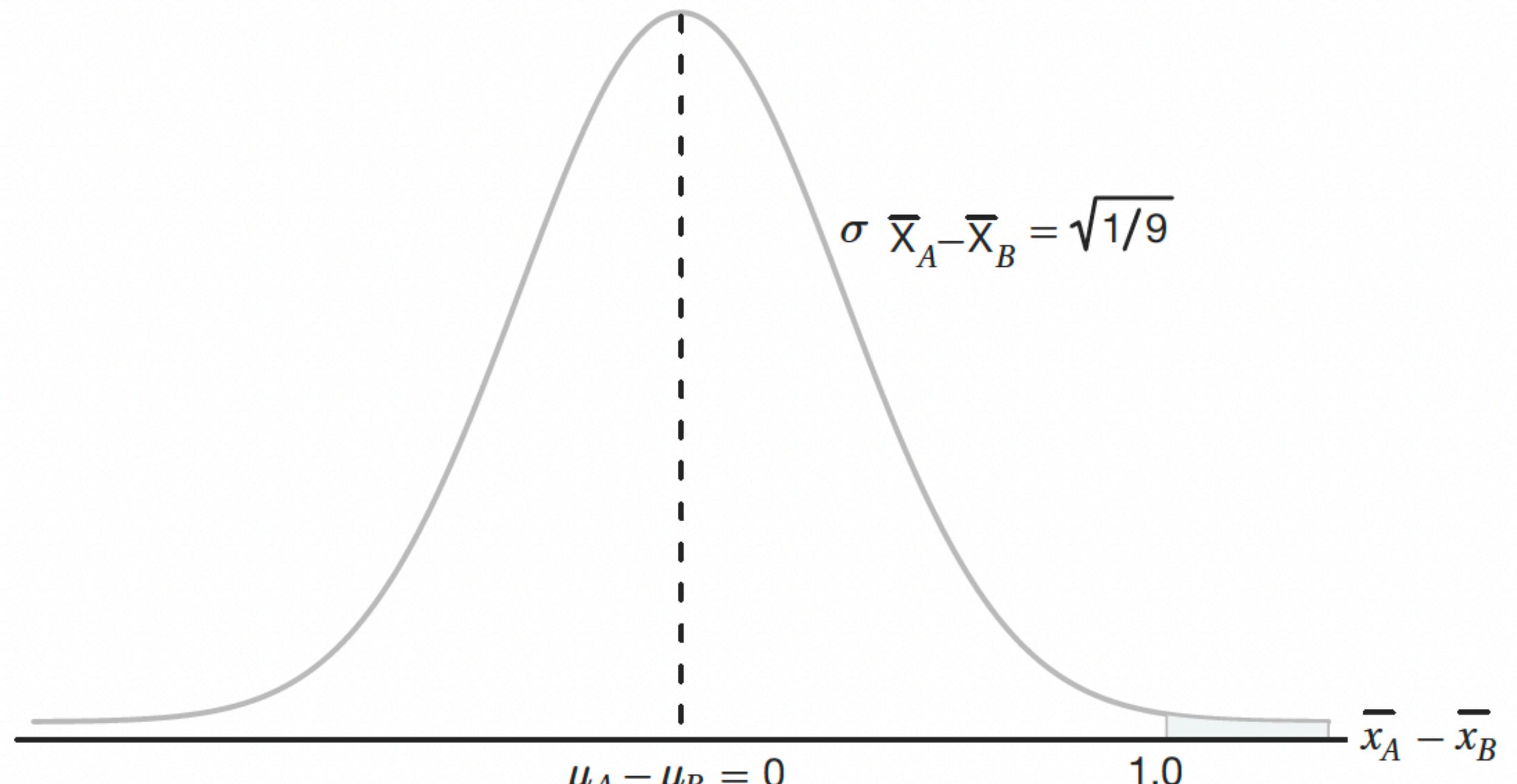


Figura 8.5: Área para el estudio de caso 8.2.

$$P(Z > 3.0) = 1 - P(Z < 3.0) = 1 - 0.9987 = 0.0013.$$

# Prueba de hipótesis

- El ejemplo anterior permite conocer la probabilidad de que el tiempo promedio de las dos muestras difiera en más de una hora.
- La conclusión es que es muy improbable encontrar esa diferencia en las muestras dado que las muestras vengan de la misma distribución.
- En la práctica, la mayoría de las veces no sabemos cuál es la distribución de la población, sino que más bien podemos calcular las medias muestrales a partir de las observaciones.
- Esto se usa para determinar **qué tan probable es que dos muestras vengan de la misma distribución**, por ejemplo cuando los tiempos de secado son idénticos para ambas pinturas.

# Distribución t

- Para realizar pruebas de hipótesis se suele utilizar la **distribución t**, que sigue a continuación.

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

- Típicamente no se suele conocer la desviación estándar de la población y por lo tanto **se utiliza la desviación estándar de la muestra, S**

# Distribución t

- La distribución t se puede utilizar:
  - Para verificar si una muestra tiene la misma distribución que una población conocida (para una muestra).
  - Para comprobar si dos muestras vienen de poblaciones distintas (para dos muestras).
- El uso de la distribución t requiere que las variables  $X_1, X_2, \dots, X_n$  que conforman las muestras sean normales
- Aquí nos enfocaremos en la comparación de dos muestras, que es ampliamente utilizada en la práctica.

# Grados de libertad

- El concepto de grados de libertad se refiere a la cantidad de información independiente en un experimento estadístico.
- Por ejemplo, cuando obtenemos una media muestral, utilizamos  $n$  muestras independientes y por lo tanto tenemos  $df = n$  grados de libertad
- Cuando calculamos la varianza, primero tenemos que obtener la media muestral, por lo que “perdemos” un grado de libertad: grados de libertad  $df = n - 1$
- Para la prueba t (varianzas iguales), los grados de libertad son  $df = n_1 + n_2 - 2$

# Distribución t

- Para dos muestras, la distribución t tiene la forma:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{s_p \sqrt{1/n_1 + 1/n_2}},$$

$$s_p^2 = \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}$$

- Donde  $s_1$  y  $s_2$  son las desviaciones estándar de la muestra
- Donde  $d_0$  es la diferencia hipotética entre las medias poblacionales (típicamente 0)

# Hipótesis

- Cuando realizamos pruebas de hipótesis, obtenemos la probabilidad  $p$  de que las dos muestras vengan de la misma distribución. Si  $\mu_1$  y  $\mu_2$  son las medias de las muestras 1 y 2, respectivamente,
- **Hipótesis nula ( $H_0$ )**: las muestras tienen la misma media
- **Hipótesis alternativa ( $H_1$ )**: las muestras no tienen la misma media

# Umbral $\alpha$

- Para realizar los análisis estadísticos se define un **umbral** dado por  $\alpha$
- Usualmente  $\alpha$  se define en 0.05, 0.01 o 0.001, dependiendo de la aplicación.
- Esto nos dice qué tan preciso es el test.
- Interesa saber si la probabilidad de las observaciones que se realizaron es menor que este  $\alpha$

# Prueba de hipótesis

## Problema

- Queremos saber si el tipo de desayuno (gallo pinto con huevo vs. corn flakes) tiene un efecto en la estatura de los costarricenses. Tenemos una muestra de estaturas de 8 personas que han comido gallo pinto y 8 personas que han comido corn flakes (mismo número de mujeres y hombres). Las estaturas del primero grupo son (en cm): 160,180,172,188,158,174,181,172. Las estaturas del segundo grupo son: 171,154,152,161,175,160,167,172. Asumimos que las poblaciones de las muestras tienen la misma varianza.
- Para probar la hipótesis de que el desayuno tiene un efecto en la estatura, asumimos que las muestras vienen de la misma distribución (misma media) y calculamos la probabilidad de obtener la diferencia de las medias observada. Usamos  $\alpha = 0.05$

# Prueba de hipótesis

## Solución

- Calculamos:

$$\bullet \bar{x}_1 = \frac{160 + 180 + 172 + 188 + 158 + 174 + 181 + 172}{8} = 173.125$$

$$\bullet \bar{x}_2 = \frac{171 + 154 + 152 + 161 + 175 + 160 + 167 + 172}{8} = 164$$

$$\bullet s_1^2 = \frac{1}{n_1 - 1} \sum (x_i - \bar{x}_1)^2 = 104.9821, s_1 = 10.245$$

$$\bullet s_2^2 = \frac{1}{n_2 - 1} \sum (x_i - \bar{x}_2)^2 = 73.1428, s_2 = 8.552$$

# Prueba de hipótesis

## Solución

- Calculamos:

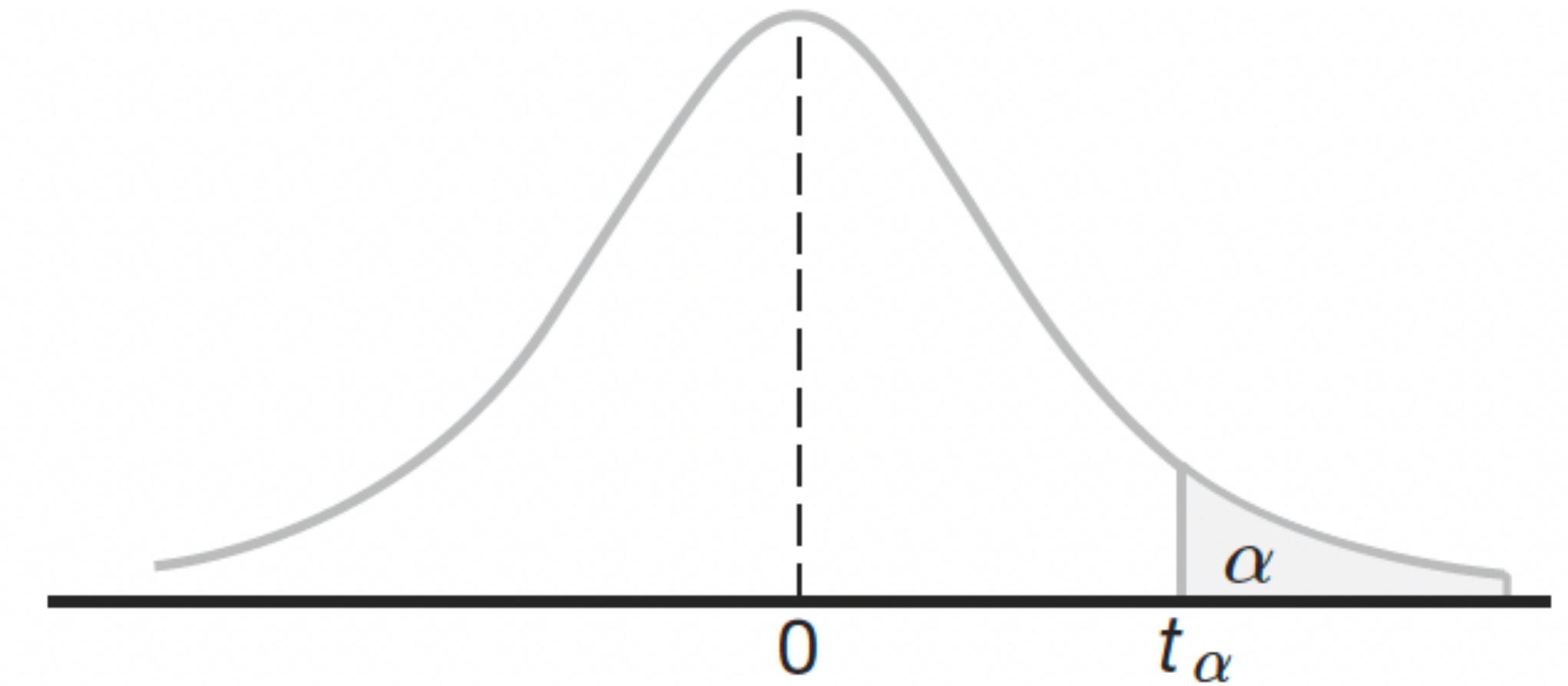
$$\bullet \quad s_p^2 = \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2} = 89.06, s_p = 9.437$$

$$\bullet \quad t = \frac{\bar{x}_1 - \bar{x}_2 - 0}{s_p \sqrt{1/n_1 + 1/n_2}} = \frac{173.125 - 164}{9.437 \sqrt{1/8 + 1/8}}$$

$$\bullet \quad t = \frac{9.125}{9.437/2} = 1.934$$

# ¿Cómo interpretamos ese valor obtenido?

- Tenemos que contrastar el valor obtenido con la distribución t, la cual se encuentra en el apéndice A.4 del libro de Walpole o al final de esta presentación.
- Para  $n_1 + n_2 - 2 = 14$  grados de libertad y una tolerancia de error de  $\alpha = 0.05$ , el valor de la tabla es 1.761
- Obtuvimos un valor mayor ( $1.934 > 1.761$ ), lo cual implica que estamos en la parte de la distribución normal a la derecha:



# ¿Cómo interpretamos ese valor obtenido?

- Concluimos entonces que, para las muestras observadas, la probabilidad de que las medias de las dos poblaciones sean iguales, es menor que 0.05
- Por lo tanto, es muy improbable que las muestras vengan de la misma distribución y entonces concluimos que **el tipo de desayuno sí influye en la estatura.**
- Comparando las muestras podemos decir que las personas que comen gallo pinto tienen mayor estatura que las que comen corn flakes

# Posibles errores en las pruebas de hipótesis

- Es importante mencionar que los métodos estadísticos no son 100% fiables y podemos incurrir en algunos errores, por ejemplo, si la probabilidad es menor que  $\alpha$  pero aún así la hipótesis nula ( $H_0$ ) es verdadera

	$H_0$ es verdadera	$H_0$ es falsa
No rechazar $H_0$	Decisión correcta	Error tipo II
Rechazar $H_0$	Error tipo I	Decisión correcta

# **En resumen**

## **La prueba t**

# Prueba t

Prueba  $t$  Para la hipótesis bilateral  
agrupada de  
dos muestras

$$H_0: \mu_1 = \mu_2,$$

$$H_1: \mu_1 \neq \mu_2,$$

rechazamos  $H_0$  al nivel de significancia  $\alpha$  cuando el estadístico  $t$  calculado

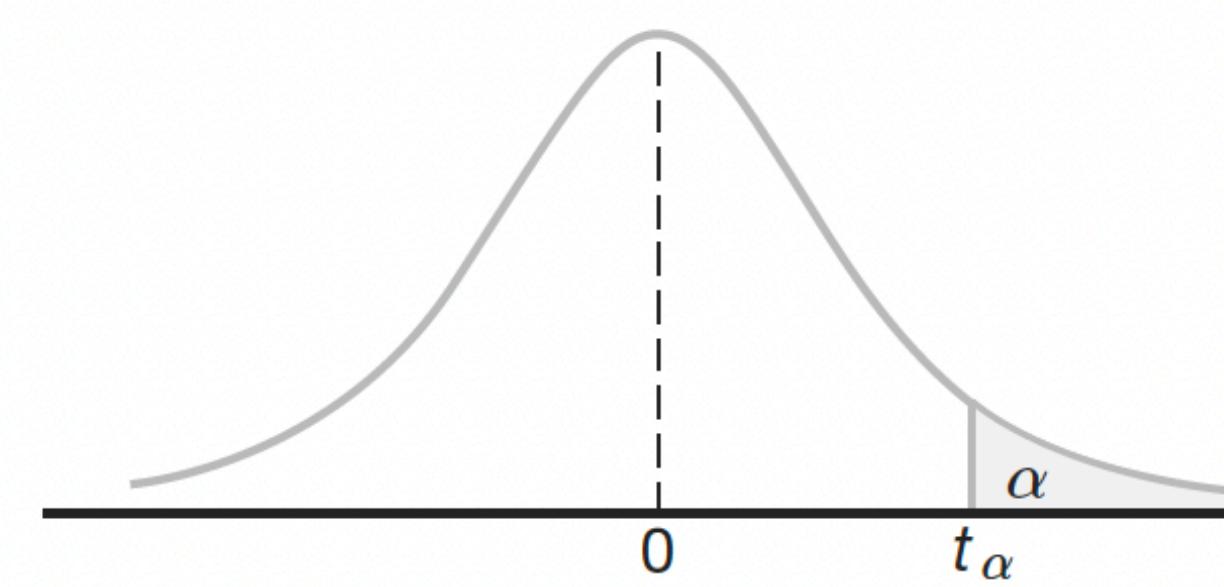
$$t = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{s_p \sqrt{1/n_1 + 1/n_2}},$$

donde

$$s_p^2 = \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}$$

excede a  $t_{\alpha/2, n_1 + n_2 - 2}$  o es menor que  $-t_{\alpha/2, n_1 + n_2 - 2}.$

---



**Tabla A.4** Valores críticos de la distribución  $t$

$v$	$\alpha$						
	<b>0.40</b>	<b>0.30</b>	<b>0.20</b>	<b>0.15</b>	<b>0.10</b>	<b>0.05</b>	<b>0.025</b>
<b>1</b>	0.325	0.727	1.376	1.963	3.078	6.314	12.706
<b>2</b>	0.289	0.617	1.061	1.386	1.886	2.920	4.303
<b>3</b>	0.277	0.584	0.978	1.250	1.638	2.353	3.182
<b>4</b>	0.271	0.569	0.941	1.190	1.533	2.132	2.776
<b>5</b>	0.267	0.559	0.920	1.156	1.476	2.015	2.571
<b>6</b>	0.265	0.553	0.906	1.134	1.440	1.943	2.447
<b>7</b>	0.263	0.549	0.896	1.119	1.415	1.895	2.365
<b>8</b>	0.262	0.546	0.889	1.108	1.397	1.860	2.306
<b>9</b>	0.261	0.543	0.883	1.100	1.383	1.833	2.262
<b>10</b>	0.260	0.542	0.879	1.093	1.372	1.812	2.228
<b>11</b>	0.260	0.540	0.876	1.088	1.363	1.796	2.201
<b>12</b>	0.259	0.539	0.873	1.083	1.356	1.782	2.179
<b>13</b>	0.259	0.538	0.870	1.079	1.350	1.771	2.160
<b>14</b>	0.258	0.537	0.868	1.076	1.345	1.761	2.145
<b>15</b>	0.258	0.536	0.866	1.074	1.341	1.753	2.131
<b>16</b>	0.258	0.535	0.865	1.071	1.337	1.746	2.120
<b>17</b>	0.257	0.534	0.863	1.069	1.333	1.740	2.110
<b>18</b>	0.257	0.534	0.862	1.067	1.330	1.734	2.101
<b>19</b>	0.257	0.533	0.861	1.066	1.328	1.729	2.093
<b>20</b>	0.257	0.533	0.860	1.064	1.325	1.725	2.086

<b>21</b>	0.257	0.532	0.859	1.063	1.323	1.721	2.080
<b>22</b>	0.256	0.532	0.858	1.061	1.321	1.717	2.074
<b>23</b>	0.256	0.532	0.858	1.060	1.319	1.714	2.069
<b>24</b>	0.256	0.531	0.857	1.059	1.318	1.711	2.064
<b>25</b>	0.256	0.531	0.856	1.058	1.316	1.708	2.060
<b>26</b>	0.256	0.531	0.856	1.058	1.315	1.706	2.056
<b>27</b>	0.256	0.531	0.855	1.057	1.314	1.703	2.052
<b>28</b>	0.256	0.530	0.855	1.056	1.313	1.701	2.048
<b>29</b>	0.256	0.530	0.854	1.055	1.311	1.699	2.045
<b>30</b>	0.256	0.530	0.854	1.055	1.310	1.697	2.042
<b>40</b>	0.255	0.529	0.851	1.050	1.303	1.684	2.021
<b>60</b>	0.254	0.527	0.848	1.045	1.296	1.671	2.000
<b>120</b>	0.254	0.526	0.845	1.041	1.289	1.658	1.980
<b><math>\infty</math></b>	0.253	0.524	0.842	1.036	1.282	1.645	1.960

**Tabla A.4** (continuación) Valores críticos de la distribución  $t$ 

$v$	$\alpha$						
	<b>0.02</b>	<b>0.015</b>	<b>0.01</b>	<b>0.0075</b>	<b>0.005</b>	<b>0.0025</b>	<b>0.0005</b>
<b>1</b>	15.894	21.205	31.821	42.433	63.656	127.321	636.578
<b>2</b>	4.849	5.643	6.965	8.073	9.925	14.089	31.600
<b>3</b>	3.482	3.896	4.541	5.047	5.841	7.453	12.924
<b>4</b>	2.999	3.298	3.747	4.088	4.604	5.598	8.610
<b>5</b>	2.757	3.003	3.365	3.634	4.032	4.773	6.869
<b>6</b>	2.612	2.829	3.143	3.372	3.707	4.317	5.959
<b>7</b>	2.517	2.715	2.998	3.203	3.499	4.029	5.408
<b>8</b>	2.449	2.634	2.896	3.085	3.355	3.833	5.041
<b>9</b>	2.398	2.574	2.821	2.998	3.250	3.690	4.781
<b>10</b>	2.359	2.527	2.764	2.932	3.169	3.581	4.587
<b>11</b>	2.328	2.491	2.718	2.879	3.106	3.497	4.437
<b>12</b>	2.303	2.461	2.681	2.836	3.055	3.428	4.318
<b>13</b>	2.282	2.436	2.650	2.801	3.012	3.372	4.221
<b>14</b>	2.264	2.415	2.624	2.771	2.977	3.326	4.140
<b>15</b>	2.249	2.397	2.602	2.746	2.947	3.286	4.073
<b>16</b>	2.235	2.382	2.583	2.724	2.921	3.252	4.015
<b>17</b>	2.224	2.368	2.567	2.706	2.898	3.222	3.965
<b>18</b>	2.214	2.356	2.552	2.689	2.878	3.197	3.922
<b>19</b>	2.205	2.346	2.539	2.674	2.861	3.174	3.883
<b>20</b>	2.197	2.336	2.528	2.661	2.845	3.153	3.850

<b>21</b>	2.189	2.328	2.518	2.649	2.831	3.135	3.819
<b>22</b>	2.183	2.320	2.508	2.639	2.819	3.119	3.792
<b>23</b>	2.177	2.313	2.500	2.629	2.807	3.104	3.768
<b>24</b>	2.172	2.307	2.492	2.620	2.797	3.091	3.745
<b>25</b>	2.167	2.301	2.485	2.612	2.787	3.078	3.725
<b>26</b>	2.162	2.296	2.479	2.605	2.779	3.067	3.707
<b>27</b>	2.158	2.291	2.473	2.598	2.771	3.057	3.689
<b>28</b>	2.154	2.286	2.467	2.592	2.763	3.047	3.674
<b>29</b>	2.150	2.282	2.462	2.586	2.756	3.038	3.660
<b>30</b>	2.147	2.278	2.457	2.581	2.750	3.030	3.646
<b>40</b>	2.123	2.250	2.423	2.542	2.704	2.971	3.551
<b>60</b>	2.099	2.223	2.390	2.504	2.660	2.915	3.460
<b>120</b>	2.076	2.196	2.358	2.468	2.617	2.860	3.373
<b><math>\infty</math></b>	2.054	2.170	2.326	2.432	2.576	2.807	3.290