# Distribuciones de muestreo importantes

Jhon Jairo Padilla Aguilar, PhD.

# Distribución chi-cuadrado

- Si $z_1, z_2, z_3, \ldots z_k$ son distribuciones normales estandarizadas, la variable aleatoria

$$x = z_1^2 + z_2^2 + \ldots + z_k^2$$

- Sigue una distribución chi-cuadrada con *k* grados de libertad

# Distribución chi-cuadrado

- La función de densidad de probabilidad chi-cuadrada es

$$f(x) = \frac{1}{2^{k/2}\Gamma\left(\frac{k}{2}\right)} x^{(k/2)-1} e^{-x/2}, x > 0$$

- Con media k y varianza 2k
- La función Gamma es:

$$\Gamma(r) = \int_0^\infty x^{r-1} e^{-x}\, dx, \quad \text{for } r > 0$$

# Distribución chi-cuadrada

- Es asimétrica (sesgada).

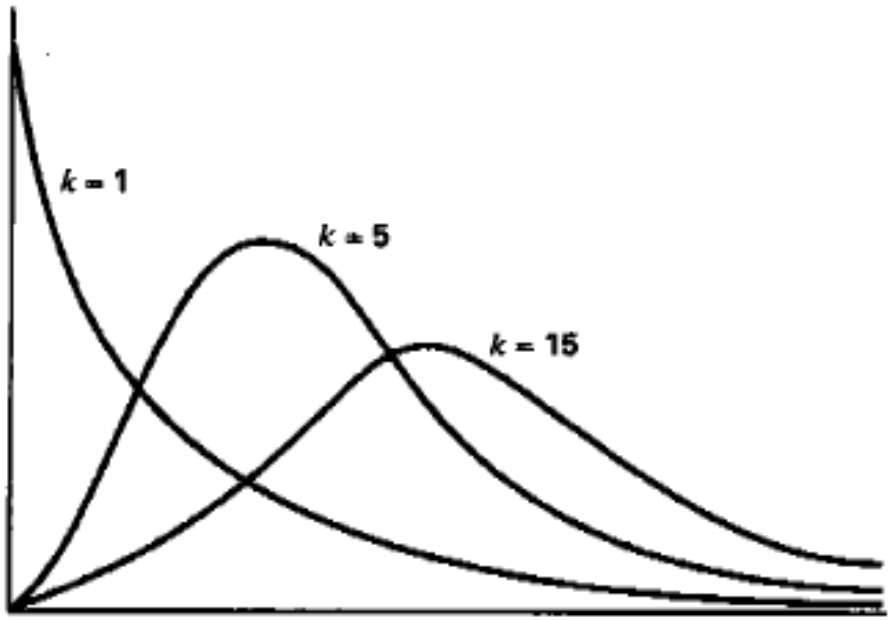



**Figura 2-6** Varias distribuciones ji-cuadrada.

# Aplicación de la distribución chi-cuadrado

- Se obtiene una distribución chi-cuadrado cuando se tienen sumas de cuadrados.
- Un ejemplo es cuando se tiene una suma como

$$\frac{SS}{\sigma^2} = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{\sigma^2} \sim \chi^2_{n-1}$$

- Esta expresión sigue una distribución chi-cuadrado con n-1 grados de libertad y se utiliza para hacer pruebas de bondad de ajuste.

# Aplicación de la distribución chi-cuadrado

- La varianza muestral puede escribirse como

$$S^2 = \frac{SS}{n-1}$$

- Si se tienen observaciones con una distribución normal con media $\mu$ y varianza $\sigma^2$, entonces la distribución de

$$S^2 \text{ es } [\sigma^2/(n-1)]\chi^2_{n-1}$$

- Por tanto, la distribución de muestreo de la varianza muestral es una constante multiplicada por la distribución chi cuadrada, si la población tiene una distribución normal.

# La distribución t

# Motivación

- El *teorema del límite central* puede aplicarse siempre y cuando el número de observaciones sea mayor o igual a 30.
- Se puede usar la *distribución normal* para este contexto, suponiendo que se conoce la desviación estándar de la población.
  - Esto requiere un buen conocimiento de proceso!!
- En muchas ocasiones no se tienen suficientes observaciones, ni se conoce la desviación estándar poblacional. Por tanto, **no puede usarse la distribución normal**.

# Solución: la distribución t

- El estadístico a utilizar es

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

- Se utiliza la desviación estándar muestral ya que no se conoce la desviación estándar poblacional
- Si el valor de n es pequeño, los valores de S fluctúan considerablemente entre muestras diferentes
- La distribución de T se desvía considerablemente de una distribución normal estándar.

# Distribución t

- Sea $X_1, X_2, X_3, \ldots, X_n$ una muestra aleatoria para una distribución normal con media desconocida y varianza desconocida. La cantidad:

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

- Tiene una distribución *t* con n-1 grados de libertad.
- La función de densidad de probabilidad t es:

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

# Distribución t

- Se publicó originalmente en 1908 en un artículo de W.S. Gosset, quien era empleado de una cervecería que no le permitía la publicación de investigaciones a sus empleados.
- Para evadir esa prohibición, Gosset publicó su trabajo bajo el nombre "Student".
- Por tanto, a la distribución *t* también se le conoce como *t de Student*.

# Relación de chi cuadrado con la distribución t

$$t_k = \frac{z}{\sqrt{\chi_k^2 / k}}$$

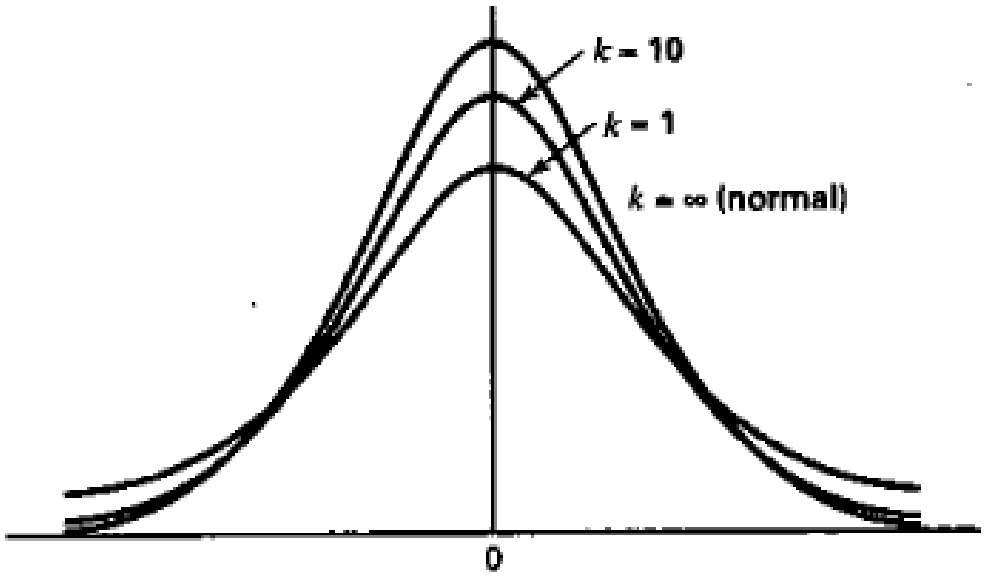



**Figura 2-7** Varias distribuciones *t*.

# A qué se parece la distribución t?

- Es simétrica alrededor de una media cero, como la distribución normal estándar.
- Ambas tienen forma de campana
- La distribución t es más variable (la probabilidad de las colas es mayor) que la normal estándar.
- La varianza de la distribución t depende del tamaño de la muestra (n) y siempre es mayor que 1.
- La distribución t tiende a una distribución normal estándar cuando n tiende a infinito.

# Distribución F

- Si tenemos dos distribuciones chi cuadrado con u y v grados de libertad respectivamente, entonces el cociente

$$F_{u,v} = \frac{\chi_u^2 / u}{\chi_v^2 / v}$$

- Sigue la distribución F con u grados de libertad en el numerador y v grados de libertad en el denominador

# Distribución F

- La distribución de probabilidad de x será

$$h(x)=\frac{\Gamma\left(\frac{u+v}{2}\right)\left(\frac{u}{v}\right)^{u/2}x^{(u/2)-1}}{\Gamma\left(\frac{u}{2}\right)\Gamma\left(\frac{v}{2}\right)\left[\left(\frac{u}{v}\right)x+1\right]^{(u+v)/2}} \qquad 0<x<\infty$$

# Distribución F

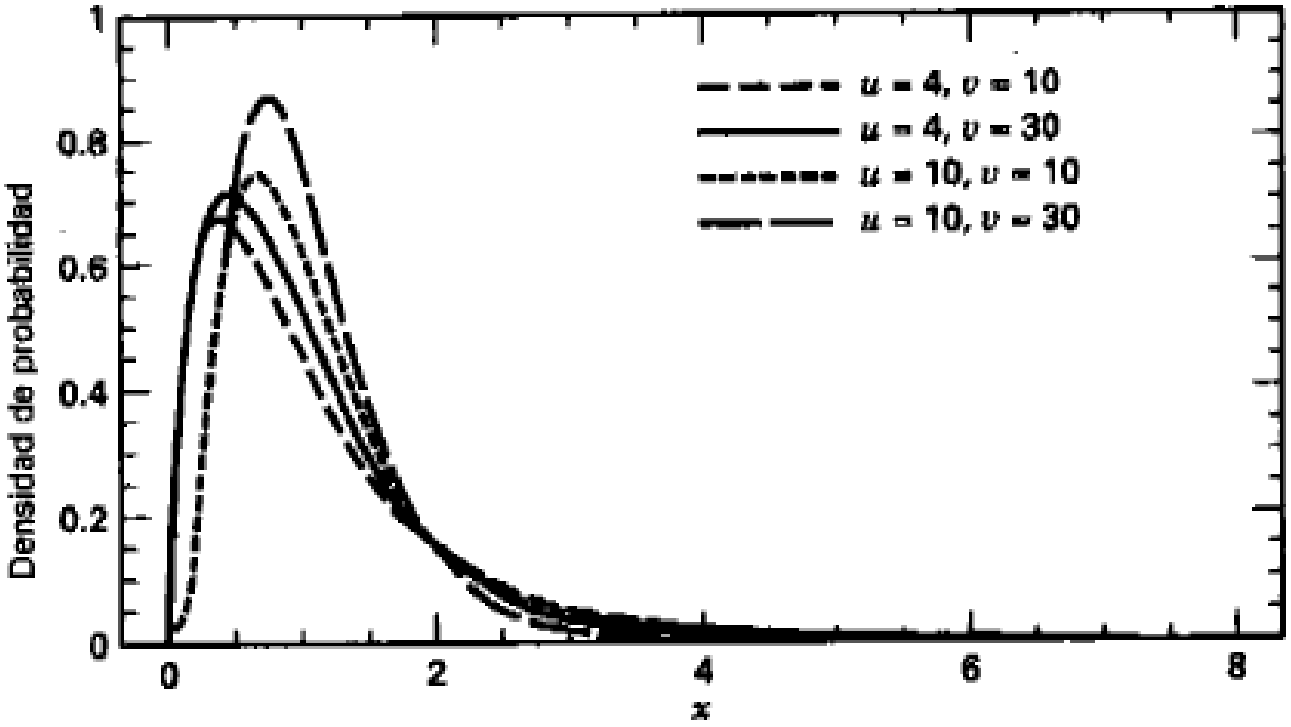



**Figura 2-8** Varias distribuciones *F*.

# Aplicación de la distribución F

- Se utiliza en el diseño de experimentos
- Ejemplo:
  - Suponga que se tienen dos poblaciones normales independientes con la misma varianza.
  - Si la primera muestra tiene $n_1$ observaciones y la segunda muestra $n_2$ observaciones, entonces el cociente

$$\frac{S_1^2}{S_2^2} \sim F_{n_1-1,\, n_2-1}$$

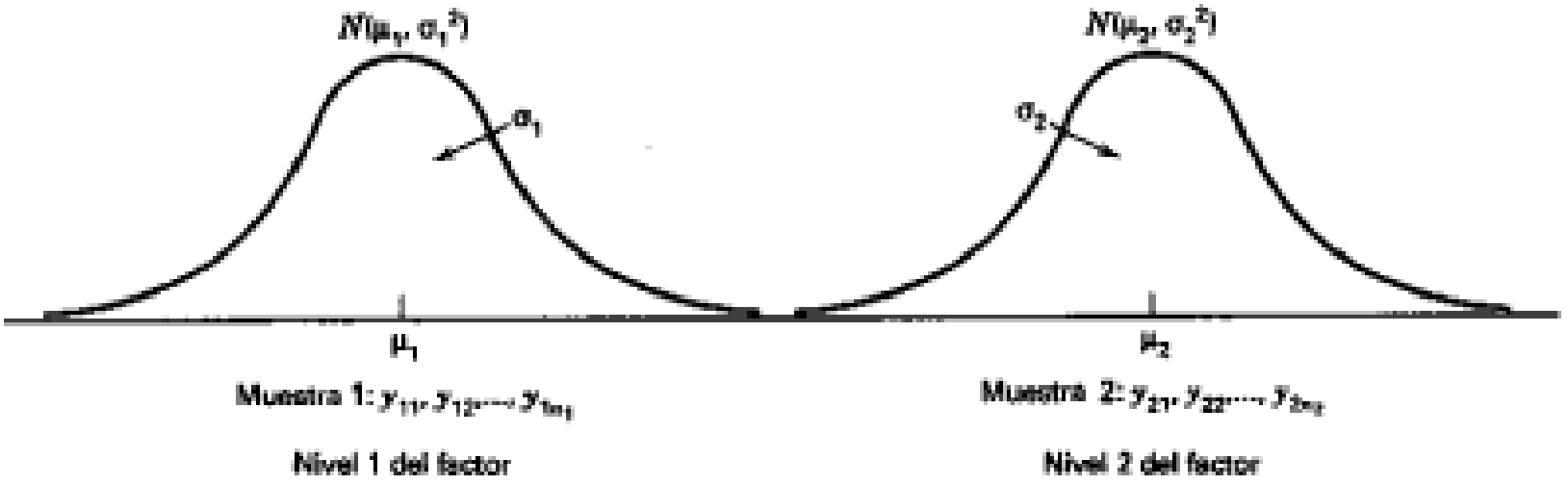
  - Donde $S_1$ y $S_2$ son las desviaciones estándar de las muestras, el cociente tenderá a una distribución F

# Inferencias acerca de las diferencias en las medias (Diseños Aleatorizados)

- Se busca determinar si dos métodos difieren en la media o si tienen la misma media.
- Se utiliza la prueba de hipótesis y los intervalos de confianza para comparar las medias de los tratamientos
- Se supone un diseño experimental completamente aleatorizado
- Se considera que los datos fueron tomados de una muestra aleatoria de una distribución normal

# La situación

- Se han tomado $n_1$ observaciones del primer factor y $n_2$ observaciones del segundo factor
- Las muestras se tomaron al azar




# Hipótesis nula y alternativa

$$H_o : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

# Prueba t de dos muestras

- Suponiendo que las varianzas de las dos muestras son iguales, se puede utilizar como estadístico de la prueba el siguiente:

$$t_0 = \frac{\bar{y}_1 - \bar{y}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

- Este estadístico permite comparar las diferencias entre las medias de los dos tratamientos
- Las $y_i$ son las medias de los tratamientos y las $n_i$ son los tamaños de las muestras.
- $S_p$ es una varianza calculada a partir de las varianzas de cada muestra así:

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

# Prueba t de dos muestras

- Para determinar si deberá rechazarse la hipótesis nula, debe compararse $t_o$ con la distribución t con $n_1+n_2-2$ grados de libertad.
- Si $|t_o| > t_{\alpha/2,n_1+n_2-2}$ , donde $t_{\alpha/2,n_1+n_2-2}$

es el punto porcentual $\alpha/2$ superior de la distribución t con $n_1+n_2-2$ grados de libertad, entonces se rechazaría $H_o$ (entonces las medias difieren)

# Justificación del procedimiento

- Se supone que el muestreo se hace de distribuciones normales independientes
- Por tanto, la distribución de la diferencia de las medias es una distribución normal

$$N\left[\mu_1 - \mu_2, \sigma^2(1/n_1 + 1/n_2)\right]$$

- Por tanto, si se conoce la desviación estándar y las medias fueran iguales, obtendríamos una distribución normal estándar (media cero y desviación estándar 1)
- Al sustituir la desv. Estandar por $S_p$, se cambia la distribución normal estándar por una distribución *t* con $n_1+n_2-2$ grados de libertad.

# Procedimiento general de la prueba

- Si $H_o$ es verdadera, $t_o$ tendría la distribución de probabilidad $t_{n1+n2-2}$ y, por consiguiente, se esperaría que 100(1-$\alpha$) por ciento de los valores de $t_o$ estén entre $-t_{\alpha/2,n_1+n_2-2}$ y $t_{\alpha/2,n_1+n_2-2}$
- Una muestra que produjera un valor de $t_0$ que estuviera fuera de estos límites sería inusual si la hipótesis nula fuera verdadera y es evidencia de que Ho deberá rechazarse
- Por tanto, la distribución t con $n_1+n_2-2$ grados de libertad es la distribución de referencia apropiada para el estadístico de prueba $t_o$ (describe adecuadamente el comportamiento de $t_o$ cuando la hipótesis nula es verdadera).

# Prueba con Hipótesis de una cola

- En algunos problemas quizá quiera rechazarse Ho únicamente si una de las medias es mayor que la otra. Por lo tanto, se especificaría una hipótesis alternativa de una cola (mayor que) y $H_o$ sólo se rechazaría si $t_o$ mayor que el estadístico. En caso de una $H_1$ de una cola con menor que, se rechazaría $H_o$ si $t_o$ menor que el estadístico.

# Ejemplo

- Suponga que se tienen los datos de fuerza de adhesión de dos muestras de cemento portland y se encuentra que

| Mortero modificado | Mortero sin modificar |
|---|---|
| $\bar{y}_1 = 16.76$ kgf / cm$^2$ | $\bar{y}_2 = 17.92$ kgf / cm$^2$ |
| $S_1^2 = 0.100$ | $S_2^2 = 0.061$ |
| $S_1 = 0.316$ | $S_2 = 0.247$ |
| $n_1 = 10$ | $n_2 = 10$ |

- Puesto que las desviaciones estándar muestrales son cercanas, podría asumirse que las desviaciones estándar poblacionales son iguales
- Se proponen las hipótesis:

$$H_o : \mu_1 = \mu_2$$

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

# Ejemplo

$$n_1 + n_2 - 2 = 10 + 10 - 2 = 18$$

$$\alpha = 0.05$$

- Se rechazaría Ho si $t_0 > t_{0.025,\,18} = 2.101$ si $t_0 < -t_{0.025,\,18} = -2.101$

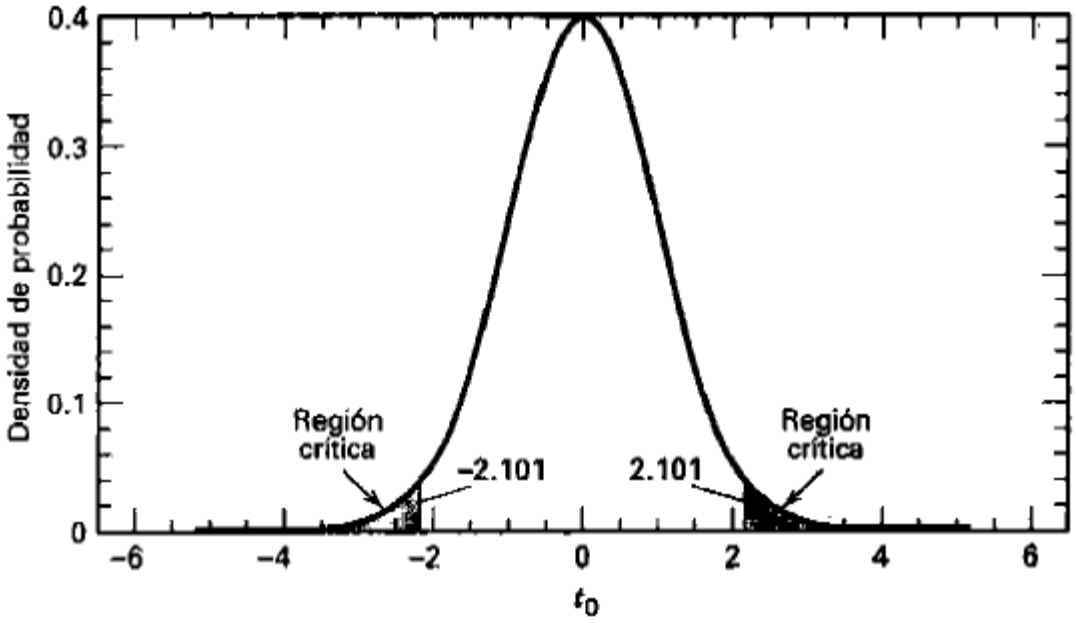



**Figura 2-10** La distribución $t$ con 18 grados de libertad con la región crítica $\pm t_{0.025,\,18} = \pm 2.101$.

# Ejemplo

$$S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$$
$$= \frac{9(0.100)+9(0.061)}{10+10-2}$$
$$= 0.081$$
$$S_p = 0.284$$

$$t_0 = \frac{\bar{y}_1-\bar{y}_2}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$
$$= \frac{16.76-17.92}{0.284\sqrt{\frac{1}{10}+\frac{1}{10}}}$$
$$= -9.13$$

Puesto que $t_0 = -9.13 < -t_{0.025,\,18} = -2.101$, se rechaza Ho y se concluye que las medias de las dos muestras son diferentes.

# El uso de valores P en la prueba de hipótesis

- Una manera de reportar los resultados de una prueba de hipótesis es estableciendo que la hipótesis nula fue rechazada o no para un valor de a o nivel de significación ($\alpha$) específico.
- Esta enunciación de las conclusiones no ofrece al responsable de la toma de decisiones una idea de si el valor calculado del estadístico de prueba apenas rebasó la región de rechazo o si se adentró bastante en la misma.
- Además, un valor de $\alpha$=0,05 podría ser riesgoso para algunos responsables de toma de decisiones.

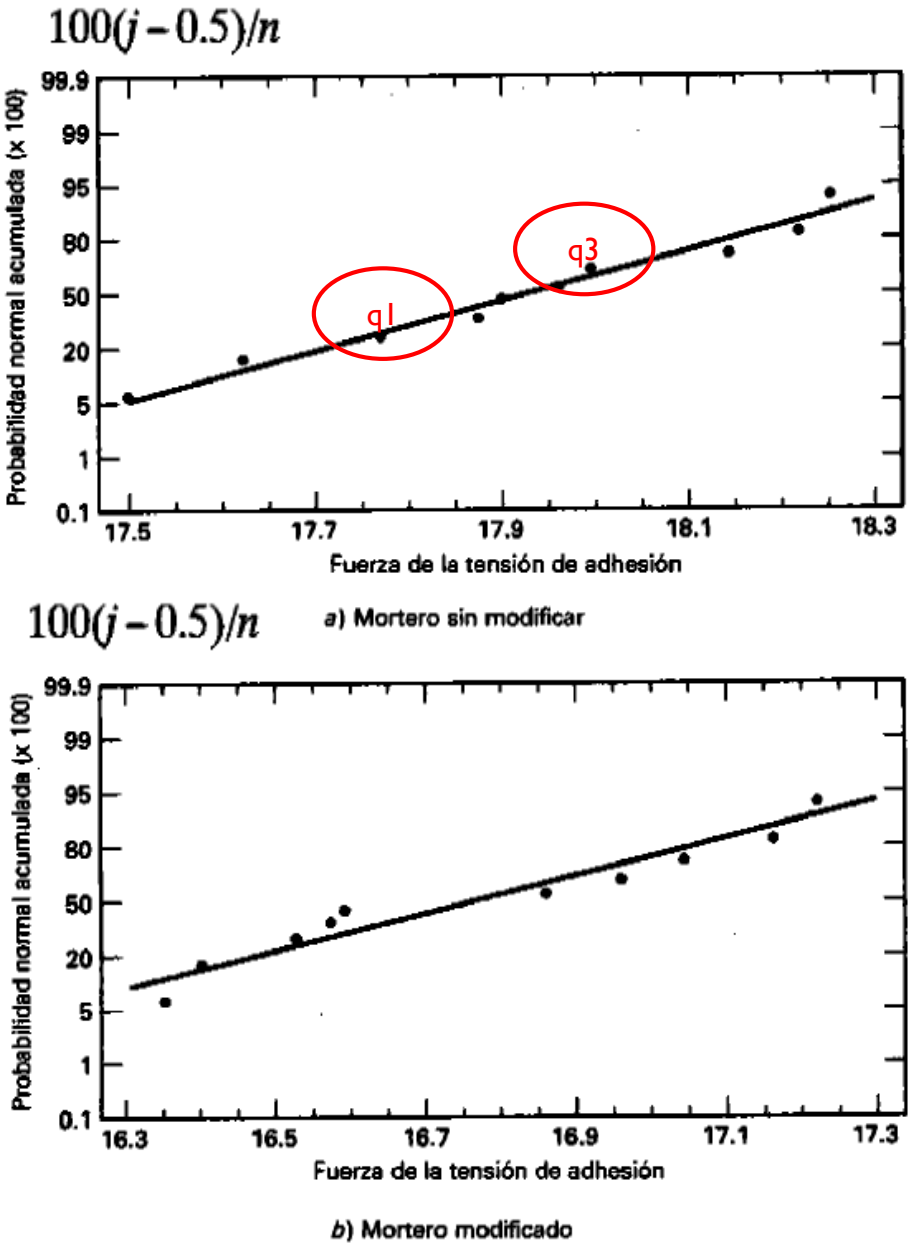
# Enfoque del valor P

- El valor de P se define como el nivel de significación menor que llevaría a rechazar la hipótesis nula Ho.
- Un valor P transmite mucha información acerca del peso de la evidencia en contra de Ho
- El responsable de la toma de decisiones puede llegar a una conclusión con cualquier nivel de significación especificado.

# Enfoque del valor P

- Se acostumbra decir: "el estadístico de prueba (y los datos) es significativo cuando se rechaza la hipótesis nula"
- El valor P puede considerarse como el menor nivel de $\alpha$ en el que los datos son significativos.
- Una vez que se conoce el valor P, el responsable de la toma de decisiones puede determinar la medida en que los datos son significativos sin que el analista de los datos imponga formalmente un nivel de significación preseleccionado.
- No es fácil calcular el valor P. Se requieren programas de computador (Muchos programas de análisis estadístico modernos lo hacen).

# Verificación de los supuestos de la prueba t




Si los puntos están cerca a la recta se considera buena aproximación la distribución normal de los datos.

# Resumen de pruebas para diferencias de medias con diseños aleatorizados

**Tabla 2-3** Pruebas para medias con varianza conocida

| Hipótesis | Estadístico de prueba | Criterios de rechazo |
|---|---|---|
| $H_0: \mu = \mu_0$<br>$H_1: \mu \neq \mu_0$ | $Z_0 = \dfrac{\bar{y} - \mu_0}{\sigma / \sqrt{n}}$ | $\lvert Z_0 \rvert > Z_{\alpha/2}$ |
| $H_0: \mu = \mu_0$<br>$H_1: \mu < \mu_0$ | | $Z_0 < -Z_\alpha$ |
| $H_0: \mu = \mu_0$<br>$H_1: \mu > \mu_0$ | | $Z_0 > Z_\alpha$ |
| $H_0: \mu_1 = \mu_2$<br>$H_1: \mu_1 \neq \mu_2$ | $Z_0 = \dfrac{\bar{y}_1 - \bar{y}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$ | $\lvert Z_0 \rvert > Z_{\alpha/2}$ |
| $H_0: \mu_1 = \mu_2$<br>$H_1: \mu_1 < \mu_2$ | | $Z_0 < -Z_\alpha$ |
| $H_0: \mu_1 = \mu_2$<br>$H_1: \mu_1 > \mu_2$ | | $Z_0 > Z_\alpha$ |

**Tabla 2-4** Pruebas para medias de distribuciones normales, varianza desconocida

| Hipótesis | Estadístico de prueba | Criterios de rechazo |
|---|---|---|
| $H_0: \mu = \mu_0$<br>$H_1: \mu \neq \mu_0$ | $t_0 = \dfrac{\bar{y} - \mu_0}{S / \sqrt{n}}$ | $\lvert t_0 \rvert > t_{\alpha/2,n-1}$ |
| $H_0: \mu = \mu_0$<br>$H_1: \mu < \mu_0$ | | $t_0 < -t_{\alpha,n-1}$ |
| $H_0: \mu = \mu_0$<br>$H_1: \mu > \mu_0$ | | $t_0 > t_{\alpha,n-1}$ |
| | **si $\sigma_1^2 = \sigma_2^2$** | |
| $H_0: \mu_1 = \mu_2$<br>$H_1: \mu_1 \neq \mu_2$ | $t_0 = \dfrac{\bar{y}_1 - \bar{y}_2}{S_p \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$<br>$v = n_1 + n_2 - 2$ | $\lvert t_0 \rvert > t_{\alpha/2,v}$ |
| | **si $\sigma_1^2 \neq \sigma_2^2$** | |
| $H_0: \mu_1 = \mu_2$<br>$H_1: \mu_1 < \mu_2$ | $t_0 = \dfrac{\bar{y}_1 - \bar{y}_2}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$ | $t_0 < -t_{\alpha,v}$ |
| $H_0: \mu_1 = \mu_2$<br>$H_1: \mu_1 > \mu_2$ | $v = \dfrac{\left(\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}\right)^2}{\dfrac{(S_1^2/n_1)^2}{n_1 - 1} + \dfrac{(S_2^2/n_2)^2}{n_2 - 1}}$ | $t_0 > t_{\alpha,v}$ |

# Una justificación alternativa de la prueba t

- La prueba t depende del supuesto fundamental de que las dos poblaciones de las que se seleccionaron las muestras al azar son normales.
- Sin embargo, las desviaciones moderadas de la normalidad no afectarán seriamente los resultados.

# Por qué?

- Si se utiliza diseño aleatorizado, se puede probar la hipótesis sin ningún supuesto respecto a la forma de la distribución (Box, Hunter y Hunter)
- Razonamiento:
  - Si los tratamientos no tienen ningún efecto, todas las 184,756 formas posibles en que podrían ocurrir las 20 observaciones son igualmente posibles
  - Hay un valor de to para cada uno de estos 184,756 arreglos.
  - Si el valor de to que se obtiene de los datos es inusualmente grande o inusualmente pequeño con referencia a los 184,756 valores, es una indicación de que las medias de las muestras son diferentes.
  - A este procedimiento se le llama **prueba de aleatorización**.

# Prueba t como prueba de aleatorización

- Puede demostrarse que la prueba t es una buena aproximación de la prueba de aleatorización
- Por tanto, usaremos las pruebas t sin prestar demasiada atención al supuesto de normalidad
- Esta es una razón simple por la que la verificación con las gráficas de probabilidad normal es un procedimiento adecuado para verificar el supuesto de normalidad.

# Elección del tamaño de la muestra

- Es uno de los aspectos más importantes de cualquier problema de diseño experimental.
- La elección del tamaño de la muestra y la probabilidad de error de tipo II ($\beta$) guardan una estrecha relación.
- En general, la probabilidad $\beta$ disminuye cuando aumenta el tamaño de la muestra.