Distribuciones Fundamentales de Muestreo

UCR – ECCI
CI-0115 Probabilidad y Estadística
Prof. Kryscia Daviana Ramírez Benavides

Distribuciones Muestrales

- La distribución de probabilidad de un estadístico se llama **distribución muestral**.
- Esta distribución depende del tamaño de la población, el tamaño de las muestras y el método de elección de las muestras.
- Existen distribuciones muestrales de $\overline{X}$ y S^2, que son el mecanismo a partir del cual se hace inferencias de los parámetros μ y σ^2.

Distribuciones Muestrales (cont.)

- La distribución muestral de $\overline{X}$ con tamaño muestral n es la distribución que resulta cuando un experimento se lleva a cabo una y otra vez y resultan los diversos valores de $\overline{X}$.
  - Esta distribución muestral describe la variabilidad de los promedios muestrales alrededor de la media de la población μ.
- Se aplica el mismo principio en el caso de la distribución de S^2.
  - Esta distribución produce información acerca de la variabilidad de los valores de s^2 alrededor de σ^2 en experimentos que se repiten.

Distribuciones Muestrales de Medias

- Suponga que se tiene una muestra aleatoria de *n* observaciones que se toma de una población normal con media μ y varianza σ^2.
- Cada observación X_i, $i = 1, 2, \ldots, n$, de la muestra aleatoria tendrá entonces la misma distribución normal que la población que se muestrea.

Distribuciones Muestrales de Medias (cont.)

- **Teorema.** Si $X_1, X_2, \ldots, X_n$ son variables aleatorias independientes que tienen distribuciones normales con medias $\mu_1, \mu_2, \ldots \mu_n$ y varianzas $\sigma_1^2, \sigma_2^2, \ldots \sigma_n^2$ respectivamente, entonces la variable aleatoria

$$Y = a_1 X_1 + a_2 X_2 + \ldots + a_n X_n$$

tiene una distribución normal con media

$$\mu_Y = a_1 \mu_1 + a_2 \mu_2 + \ldots + a_n \mu_n$$

y varianza

$$\sigma_Y^2 = a_1^2 \sigma_1^2 + a_2^2 \sigma_2^2 + \ldots + a_n^2 \sigma_n^2$$

Distribuciones Muestrales de Medias (cont.)

- Según el teorema donde se establece la propiedad reproductiva de la distribución normal, se concluye que

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

tiene distribución normal con media y varianza

$$\mu_{\overline{X}} = \frac{\mu + \mu + \ldots + \mu}{n} = \mu \quad \sigma^2_{\overline{X}} = \frac{\sigma^2 + \sigma^2 + \ldots + \sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Distribuciones Muestrales de Medias (cont.)

- Aunque se tomen muestras de una población con distribución desconocida, finita o infinita, la distribución muestral de $\overline{X}$ aún será aproximadamente normal con media μ y varianza σ^2/n, siempre que el tamaño de la muestra sea grande.
- **Teorema del Límite Central.** Si $\overline{X}$ es la media de una muestra aleatoria de tamaño n tomada de una población con media μ y varianza σ^2, entonces la forma límite de la distribución de

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

conforme $n \to \infty$, es la distribución normal estándar $n(z;0,1)$.

Distribuciones Muestrales de Medias (cont.)

- La aproximación normal para $\overline{X}$ por lo general será buena:
  - Si $n \geq 30$ sin importar la forma de la población.
  - Si $n < 30$, sólo si la población no es muy diferente a una distribución normal.
  - Si se sabe que la población es normal, la distribución muestral de la media seguirá una distribución normal exacta, no importa que tan pequeño sea el tamaño de las muestras.

Distribuciones Muestrales de Medias (cont.)

- Inferencias sobre la media de la población:
  - Una aplicación muy importante del teorema del límite central es la determinación de valores razonables de la media de la población μ.
  - Se utiliza para la prueba de hipótesis, estimación, control de calidad, y otros.
- Distribución muestral de la diferencia entre dos promedios:
  - Una aplicación importante de estas distribuciones incluye dos poblaciones, para compararlas.
  - Esta comparación es la diferencia de las medias de las poblaciones.

Distribuciones Muestrales de Medias (cont.)

- **Teorema.** Si se extraen al azar muestras independientes de tamaño n_1 y n_2 de dos poblaciones, discretas o continuas, con medias μ_1 y μ_2, y varianzas $\sigma^2{}_1$ y $\sigma^2{}_2$, respectivamente, entonces la distribución muestral de las diferencias de las medias, $\overline{X}_1 - \overline{X}_2$, está distribuida aproximadamente de forma normal con media y varianza dadas por

$$\mu_{\overline{X}_1-\overline{X}_2} = \mu_1 - \mu_2 \qquad \sigma^2_{\overline{X}_1-\overline{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

De aquí se obtiene Z, es aproximadamente una variable normal estándar

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

Distribuciones Muestrales de Medias (cont.)

- La aproximación normal para $\overline{X}_1 - \overline{X}_2$ por lo general será buena:
  - Si $n_1 \geq 30$ y $n_2 \geq 30$ sin importar la forma de las dos poblaciones.
  - Si $n_1 < 30$ y $n_2 < 30$, sólo si las dos poblaciones no son muy diferentes a una distribución normal.
  - Si se sabe que las dos poblaciones son normales, la distribución muestral de la diferencia de las medias seguirá una distribución normal exacta, no importa que tan pequeño sea el tamaño de las muestras.

Distribución Muestral de S^2

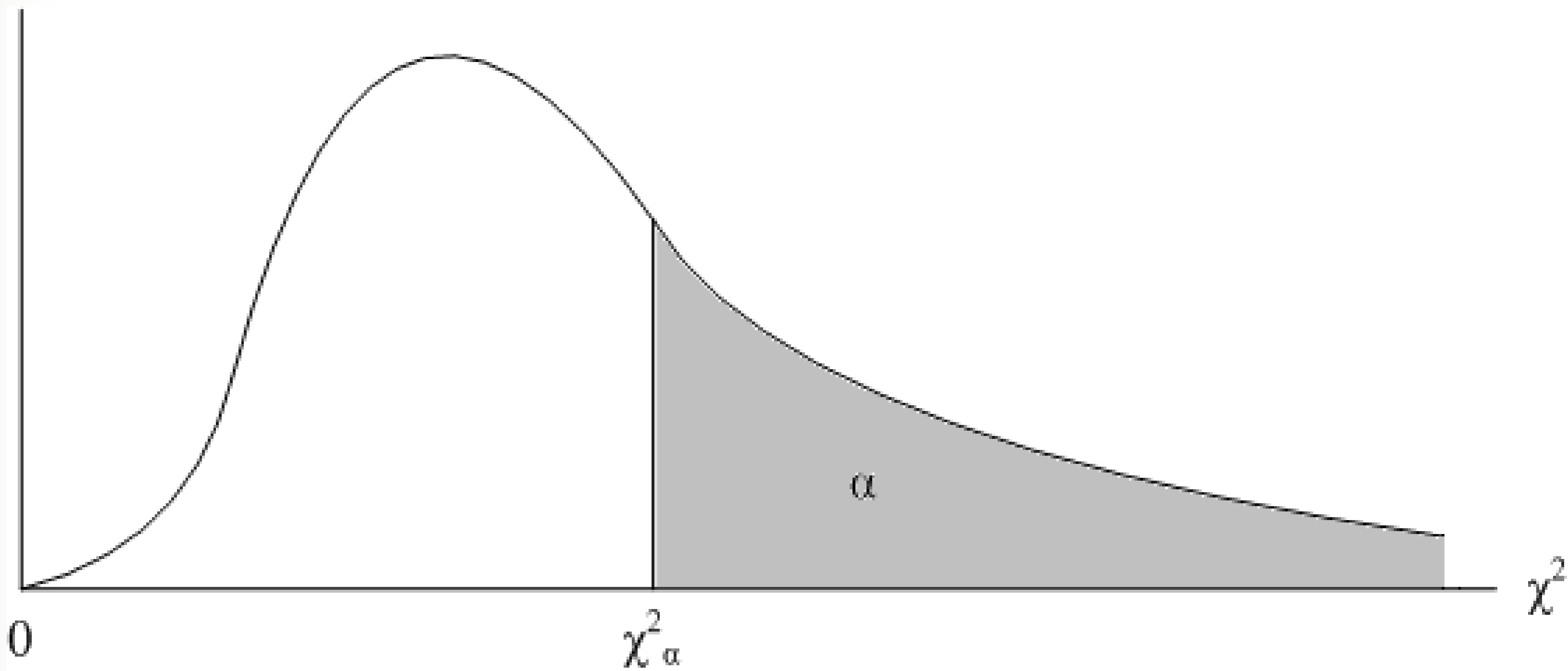
- Si S^2 es la varianza de la muestra aleatoria de tamaño n que se toma de una población normal que tiene la varianza σ^2, entonces la estadística

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \frac{\left(X_i - \overline{X}\right)^2}{\sigma^2}$$

  tiene distribución ji cuadrado con $v = n - 1$ grados de libertad.

- La tabla A.5 da los valores de χ^2_α para diversos valores de α y v. Las áreas α son los encabezados de las columnas; los grados de libertad v se dan en la columna izquierda; y las entradas de las tabla son lo valores χ^2.

Distribución Muestral de S^2 (cont.)



Distribución Muestral de S^2 (cont.)

- Exactamente 95% de una distribución ji cuadrado yace entre $\chi^2_{0.975}$ y $\chi^2_{0.025}$.
- Un valor χ^2 que cae a la derecha de $\chi^2_{0.025}$ es improbable que ocurra, a menos que el valor supuesto de σ^2 sea demasiado pequeño.
- De manera similar, un valor χ^2 que cae a la izquierda de $\chi^2_{0.975}$ es improbable que ocurra, a menos que el valor supuesto de σ^2 sea demasiado grande.
- Es decir, es posible entre un valor χ^2 a la izquierda de $\chi^2_{0.975}$ o a la derecha de $\chi^2_{0.025}$ cuando σ^2 es correcta, pero si esto debe ocurrir, es más probable que el valor supuesto de σ^2 sea un error.

Distribución Muestral de S^2 (cont.)

- Grados de libertad como medición de la información muestral:
  - Cuando los datos (los valores en la muestra) se utilizan para calcular la media, hay 1 grado de libertad menos en la información que se utiliza para estimar la varianza.

Distribución *t*

- En muchos escenarios experimentales el conocimiento de σ ciertamente no es más razonable que el conocimiento de la media de la población μ.
- A menudo una estimación de σ la debe proporcionar la misma información muestral que produce el promedio muestral $\bar{x}$.
- Como resultado, una estadística natural a considerar para tratar con las inferencias sobre μ es

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

puesto que *S* es el análogo de la muestra para σ.

Distribución *t* (cont.)

- Si el tamaño de la muestra es pequeño, los valores de S^2 fluctúan de forma considerable de una muestra a otra, y la distribución T se desvía de forma apreciable de la distribución normal estándar.
- Si el tamaño de la muestra es suficientemente grande, $n \geq 30$, la distribución T no difiere de manera considerable de la normal estándar.
- Sin embargo, si $n < 30$, es útil tratar con la distribución exacta de T.

Distribución *t* (cont.)

- Para desarrollar la distribución muestral de T se supondrá que la muestra aleatoria se seleccionó de una población normal: entonces, se puede escribir

$$T = \frac{(\overline{X} - \mu)/(\sigma/\sqrt{n})}{\sqrt{S^2/\sigma^2}} = \frac{Z}{\sqrt{V/(n-1)}}$$

  donde Z tiene distribución normal estándar y V tiene distribución ji cuadrado con $v = n - 1$ grados de libertad.

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \qquad V = \frac{(n-1)S^2}{\sigma^2}$$

- En poblaciones normales $\overline{X}$ y S^2 son independientes, y en consecuencia lo son Z y V.

Distribución t (cont.)

- **Teorema.** Sea Z una variable aleatoria normal estándar y V una variable aleatoria ji cuadrado con v grados de libertad. Si Z y V son independientes, entonces la distribución de la variable aleatoria T, donde
$$T = \frac{Z}{\sqrt{V/v}}$$
está dada por
$$h(t) = \frac{\Gamma[(v+1)/2]}{\Gamma[v/2]\sqrt{\pi v}}\left(1+\frac{t^2}{v}\right)^{-(v+1)/2} \qquad -\infty < t < +\infty$$
Esta se conoce como la **distribución *t*** con v grados de libertad, $v = n - 1$ si la muestra tiene tamaño n.

Distribución *t* (cont.)

- **Corolario.** Sean $X_1, X_2, \ldots, X_n$ variables aleatorias independientes que son normales con media μ y desviación estándar σ. Sea

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

Entonces la variable aleatoria $T = \dfrac{\overline{X} - \mu}{S/\sqrt{n}}$ tiene una

distribución *t* con $v = n - 1$ grados de libertad.

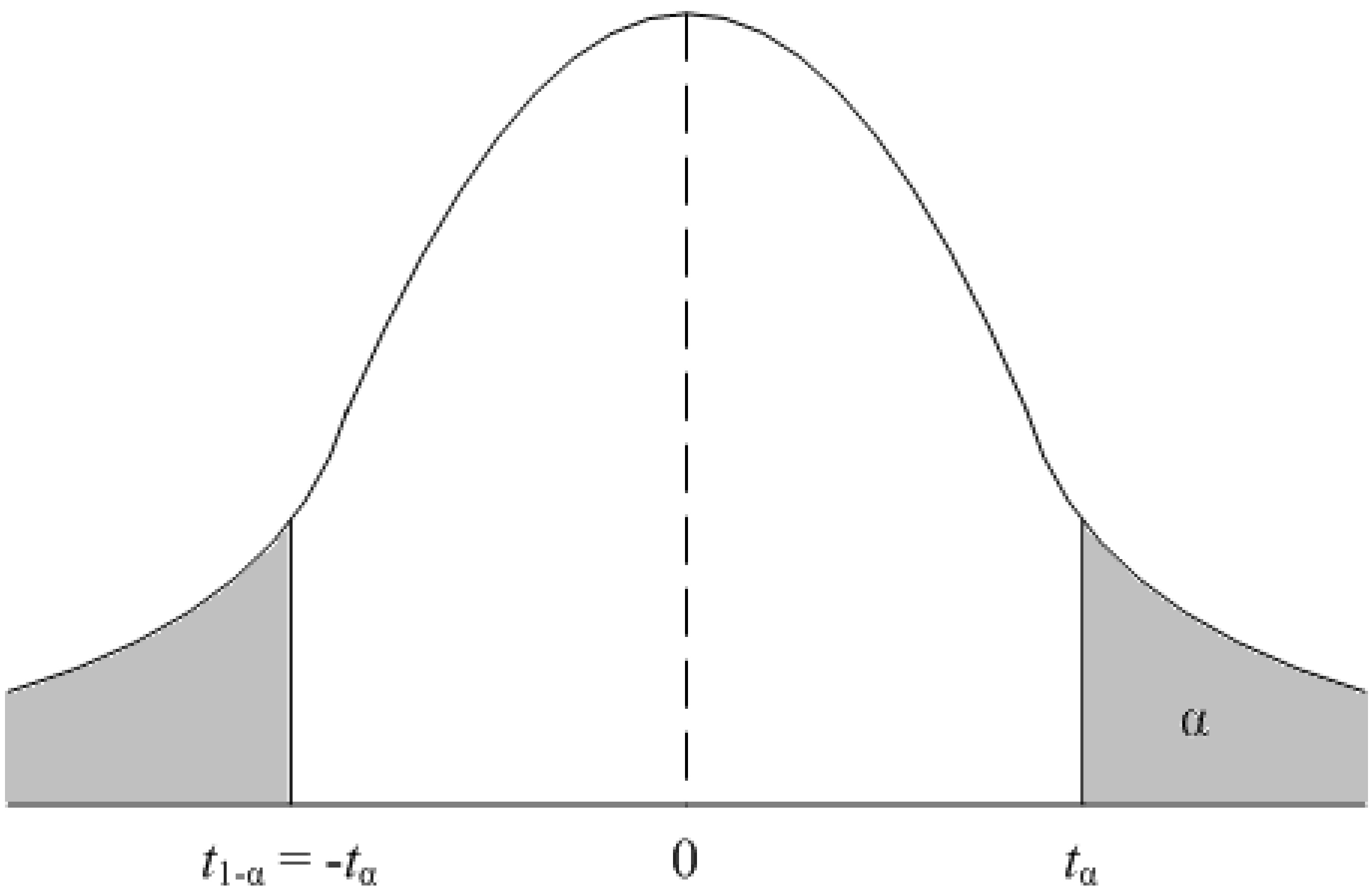
Distribución *t* (cont.)

- A la distribución *t* se le suele llamar como distribución *t* de Student.
- La distribución de T es similar a la distribución de Z, pues ambas son simétricas alrededor de una media de cero y ambas tienen forma de campana.
- La diferencia entre las dos distribuciones es que la distribución *t* es más variable que la distribución normal estándar, ya que los valores de T dependen de las fluctuaciones de $\overline{X}$ y S^2, mientras que los valores de Z dependen sólo de $\overline{X}$ de una muestra a otra.

Distribución *t* (cont.)

- La distribución de T difiere de la de Z en que la varianza de T depende del tamaño de la muestra y siempre es mayor que 1.
- Cuando el tamaño de la muestra tiende a infinito, $n \rightarrow \infty$ por lo que $v = \infty$, las dos distribuciones serán la misma.
- Se acostumbra a representar con t_α el valor t por arriba del cual se encuentra un área igual a α.
- Como la distribución t es simétrica alrededor de una media de cero, se tiene $t_{1-\alpha} = -t_\alpha$.

Distribución *t* (cont.)



Distribución *t* (cont.)

- Exactamente 95% de una distribución *t* con $v = n - 1$ grados de libertad caen entre $-t_{0.025}$ y $t_{0.025}$.
- Un valor *t* que cae por debajo de $-t_{0.025}$ o por arriba de $t_{0.025}$ tiende hacer creer que ha ocurrido un evento muy raro o quizá que la suposición acerca de μ es un error.
- Si esto ocurre, se toma la última decisión y se afirma que el valor supuesto de μ es erróneo.
- De hecho, un valor *t* que cae por debajo de $-t_{0.01}$ o por arriba de $t_{0.01}$ proporcionaría incluso fuerte evidencia de que el valor supuesto de μ es bastante improbable.

Distribución *t* (cont.)

- La distribución *t* se usa de manera extensa en problemas que tienen que ver con inferencia acerca de la media de la población o en problemas que implican muestras comparativas.
- El uso de la d distribución *t* y la consideración del tamaño de la muestra no se relacionan con el teorema del límite central.
- El uso de la distribución normal estándar en lugar de T para $n \geq 30$ sólo implica que S es un estimador suficientemente bueno de σ en este caso.

Distribución *F*

- La distribución *F* encuentra enorme aplicación en la comparación de varianzas muestrales. Las aplicaciones se encuentran en problemas que involucran dos o más muestras.
- La estadística *F* se define como la razón de dos variables aleatorias ji cuadradas independientes, dividida cada una entre su número de grados de libertad. De aquí, se puede escribir

$$F = \frac{U/v_1}{V/v_2}$$

donde *U* y *V* son variables aleatorias independientes que tienen distribuciones ji cuadradas con v_1 y v_2 grados de libertad, respectivamente.

Distribución F (cont.)

- **Teorema.** Sean U y V dos variables aleatorias independientes que tienen distribuciones ji cuadradas con v_1 y v_2 grados de libertad, respectivamente. Entonces la distribución de la variable aleatoria F, donde $F = \dfrac{U/v_1}{V/v_2}$ está dada por

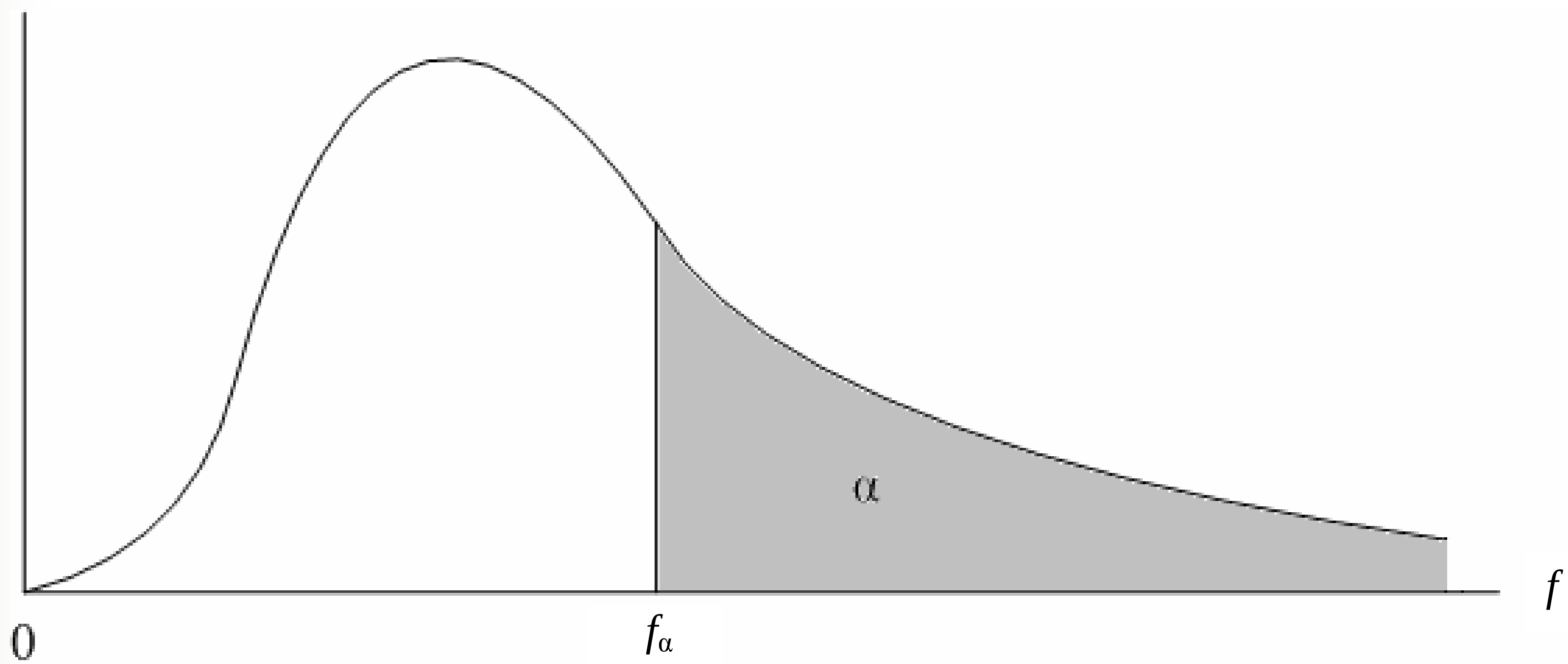
$$h(f) = \begin{cases} \dfrac{\Gamma[(v_1+v_2)/2](v_1/v_2)^{v_1/2}}{\Gamma(v_1/2)\Gamma(v_2/2)} * \dfrac{f^{v_1/2-1}}{(1+v_1 f/v_2)^{(v_1+v_2)/2}} & 0 < f < +\infty \\ 0 & \text{otro caso} \end{cases}$$

Esta se conoce como la **distribución *F*** con v_1 y v_2 grados de libertad.

Distribución *F* (cont.)

- La curva de la distribución *F* depende no sólo de los dos parámetros v_1 y v_2, sino también del orden en el que se establecen. Una vez que se dan estos dos valores, se puede identificar la curva.
- Sea f_α por arriba del cual se encuentra un área igual a α. La tabla A.6 da valores de f_α sólo para $\alpha = 0.05$ y $\alpha = 0.01$ para varias combinaciones de los grados de libertad v_1 y v_2.

Distribución *F* (cont.)



Distribución *F* (cont.)

- Por medio del siguiente teorema, la tabla A.6 también se puede utilizar para encontrar valores de $f_{0.95}$ y $f_{0.99}$.
- **Teorema.** Al escribir $f_\alpha(v_1,v_2)$ para f_α con v_1 y v_2 grados de libertad, se obtiene

$$f_{1-\alpha}(v_1, v_2) = \frac{1}{f_\alpha(v_1, v_2)}$$

Distribución F (cont.)

- **Teorema.** Si S^2_1 y S^2_2 son las varianzas de muestras aleatorias independientes de tamaño n_1 y n_2 tomadas de poblaciones normales con varianza σ^2_1 y σ^2_2, respectivamente, entonces

$$F = \frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2} = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2}$$

tiene una distribución F con $v_1 = n_1 - 1$ y $v_2 = n_2 - 1$ grados de libertad.

Distribución *F* (cont.)

- La distribución *F* se usa en situaciones de dos muestras para extraer inferencias acerca de las varianzas de población.
- También, se aplica a muchos otro tipos de problemas en los que las varianzas están involucradas.
- De hecho, la distribución *F* se llama **distribución de razón de varianzas**.

Referencias Bibliográficas

- Walpole, R.E.; Myers, R.H.; Myers, S.L. & Ye, K. "Probabilidad y estadística para ingeniería y ciencias". Octava Edición. Pearson Prentice-Hall. México, 2007.