Distribuciones Fundamentales de Muestreo y Descripciones de Datos

UCR – ECCI

CI-1352 Probabilidad y Estadística

Prof. M.Sc. Kryscia Daviana Ramírez Benavides

Muestreo Aleatorio

- En este tipo de muestreo, todos los individuos de la población pueden formar parte de la muestra, tienen una probabilidad positiva.
- El resultado de un experimento estadístico se puede registrar como un valor numérico o como una representación descriptiva.
  - Cuando se lanza un par de dados y el total es el resultado de interés, se registra un valor numérico.
  - Cuando a los estudiantes de cierta escuela se les hace pruebas de sangre y el tipo sanguíneo es de interés, se registra una representación descriptiva.
- En cualquier estudio, el número de observaciones posibles puede ser pequeño, grande pero finito o infinito.

Muestreo Aleatorio (cont.)

- Una **población** consiste en la totalidad de las observaciones en las que se está interesado.
  - Conjunto de todos los elementos que cumplen una determinada característica.
  - Conjunto de todos los valores de una variable aleatoria.
- Los elementos de la población se llaman observaciones, individuos o unidades estadísticas.
- El número de observaciones en la población se define como el **tamaño de la población**.
  - El número total de observaciones puede ser finito o infinito.

Muestreo Aleatorio (cont.)

- La **variable estadística** es una propiedad característica de la población que estamos interesados en estudiar.
- Tipos de variables estadísticas:
  - **Cualitativa:** No se expresa mediante un número. Por ejemplo, el tipo sanguíneo de los estudiantes de cierta escuela.
  - **Cuantitativa:** Se expresa mediante un número, hay dos tipos:
    - **Cuantitativa Discreta:** Sólo admite valores aislados, toma un número determinado de valores. Por ejemplo, el resultado total que se obtiene a lanzar dos dados.
    - **Cuantitativa Continua:** Puede admitir cualquier valor dentro de un intervalo, puede tomar cualquier valor entre los valores dados. Por ejemplo, medir la presión atmosférica cada día del pasado al futuro.

Muestreo Aleatorio (cont.)

- Una variable estadística cualitativa se puede convertir a una variable aleatoria discreta, para poner realizar su estudio y análisis.
- Cada observación en una población es un valor de una variable aleatoria X que tiene alguna distribución de probabilidad $f(x)$.
  - Se puede hablar de población binomial, población normal, o en general, **la población $f(x)$**, para referirse a una población cuyas observaciones son valores de una variable aleatoria que tiene una distribución binomial, una distribución normal o una distribución $f(x)$.
  - Por lo tanto, la media y la varianza de una variable aleatoria o distribución de probabilidad también se les denomina la media y la varianza de la población correspondiente.

Muestreo Aleatorio (cont.)

- En el campo de la inferencia estadística el estadístico se interesa en llegar a conclusiones con respecto a la población cuando es imposible o poco práctico observar todo el conjunto de observaciones que constituyen la población.
  - La población de una producción de cierto producto, sería imposible probar toda la producción si se tienen que vender.
  - Los costos exorbitantes también pueden ser un factor prohibitivo para estudiar toda la población.
- Por lo que se depende de un subconjunto de observaciones para hacer inferencias con respecto a la población.
- Una **muestra** es un subconjunto de una población.

Muestreo Aleatorio (cont.)

- Si se quiere inferencias válidas a partir de la muestra para la población, se debe obtener muestras que sean representativas de la población.
- Cualquier procedimiento de muestreo que produzca inferencias que sobreestimen o subestimen de forma consistente alguna característica de la población se dice que está **sesgado**.
- Para evitar cualquier posibilidad de sesgo en el procedimiento de muestreo, es deseado elegir una **muestra aleatoria** en el sentido de que las observaciones se realizan de forma independiente y al azar.

Muestreo Aleatorio (cont.)

- Sean X_1, X_2, …, X_n variables aleatorias independientes, cada una con la misma distribución de probabilidad $f(x)$. Se define entonces a X_1, X_2, …, X_n como una muestra aleatoria de tamaño n de la población $f(x)$ y se escribe su distribución de probabilidad conjunta como

$$f(x_1, x_2, \ldots, x_n) = f(x_1) f(x_2) \ldots f(x_n)$$

Algunos Estadísticos Importantes

- El propósito principal al seleccionar muestras aleatorias es obtener información acerca de los parámetros desconocidos de la población.
- Por ejemplo, se quiere saber la proporción de una población que toman una marca de café determinada.
  - Aquí se podría preguntar a cada uno de los bebedores de café de la población en cuestión, si toman la marca de café.
  - En su lugar, se selecciona una muestra aleatoria grande y se calcula la proporción $\widehat{p}$ de personas que prefieren la marcad de café.
  - El valor $\widehat{p}$ se utiliza ahora para hacer una inferencia con respecto a la proporción p verdadera.

Algunos Estadísticos Importantes (cont.)

- Ahora, $\widehat{p}$ es una función de los valores observados en la muestra aleatoria; como son posibles muchas muestras aleatorias a partir de la misma población, se espera que $\widehat{p}$ variara algo de una muestra a otra.
- Es decir, $\widehat{p}$ es un valor de una variable aleatoria que representamos con P.
- Tal variable aleatoria se llama **estadístico**, la cual se puede definir como cualquier función de las variables aleatorias que forman una muestra aleatoria.

Algunos Estadísticos Importantes – Tendencia Central de la Muestra (cont.)

- Si $X_1, X_2, \ldots, X_n$ representan una muestra aleatoria de tamaño n, entonces la **media de la muestra** se define mediante el estadístico

$$\overline{X} = \sum_{i=1}^{n} \frac{X_i}{n} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

- Si el estadístico $\overline{X}$ toma el valor $\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$

  cuando X_1 toma el valor de x_1, X_2 toma el valor de x_2, y así sucesivamente.

Algunos Estadísticos Importantes – Tendencia Central de la Muestra (cont.)

- Si $X_1, X_2, \ldots, X_n$ representan una muestra aleatoria de tamaño n, acomodada en orden creciente de magnitud, entonces la **mediana de la muestra** se define mediante el estadístico

$$\tilde{X} = \begin{cases} X_{(n+1)/2} & \text{si } n \text{ es impar} \\ \dfrac{X_{n/2} + X_{(n/2)+1}}{2} & \text{si } n \text{ es par} \end{cases}$$

Algunos Estadísticos Importantes – Tendencia Central de la Muestra (cont.)

- Si X_1, X_2, …, X_n, no necesariamente diferentes, representan una muestra aleatoria de tamaño n, entonces la **moda de la muestra** $\boldsymbol{M}$ es aquel valor de la muestra que ocurre más a menudo o con mayor frecuencia.
- La moda puede no existir, y cuando existe no necesariamente es única.

Algunos Estadísticos Importantes – Tendencia Central de la Muestra (cont.)

- La media de la muestra:
  - Es la medida de localización central más comúnmente utilizada en estadística.
  - Emplea toda la información disponible.
  - Las distribuciones de medias que se obtienen en muestreos repetidos de una población son bien conocidos, y en consecuencia los métodos que se utilizan en la inferencia estadística para estimar μ se basan en la media de la muestra.
  - La única desventaja real, es que puede resultar afectada de manera adversa por valores extremos.

Algunos Estadísticos Importantes – Tendencia Central de la Muestra (cont.)

- La mediana de la muestra:
  - Es fácil de calcular si el número de observaciones es relativamente pequeño.
  - No resulta influida por valores extremos.
  - Al tratar con muestras que se seleccionan de poblaciones, las medias de las muestras por lo general no variarán tanto de una muestra a otra como las medianas. Por lo tanto, si se desea estimar el centro de una población con base en un valor de la muestra, la media es más estable que la mediana.

Algunos Estadísticos Importantes – Tendencia Central de la Muestra (cont.)

- La moda de la muestra:
  - Es la menos utilizada de las tres.
  - Para conjuntos pequeños su valor casi no tiene utilidad, si es que existe.
  - Sólo tiene sentido significativo en una gran cantidad de datos.
  - No requiere cálculo, lo que se considera una ventaja.
  - Se puede usar para datos cualitativos como cuantitativos, lo que se considera una ventaja.

Algunos Estadísticos Importantes – Variabilidad en la Muestra (cont.)

- Las medidas de localización central o posición no dan por sí mismas una descripción adecuada de los datos. Es importante conocer cómo se dispersan las observaciones del promedio.
- La variabilidad de una muestra juega un papel muy importante en el análisis de datos.
  - La variabilidad de un proceso y de un producto es un hecho real en los sistemas científicos y de ingeniería.
  - La variabilidad en valores de población y datos de una muestra es un hecho real.

Algunos Estadísticos Importantes – Variabilidad en la Muestra (cont.)

- El **rango** (recorrido o amplitud) de una muestra aleatoria X_1, X_2, …, X_n, se define con el estadístico $X_{\max} - X_{\min}$, donde $X_{\min}$ y $X_{\max}$ son, respectivamente, las observaciones más grande y más pequeña de la muestra.
- El rango falla al medir la variabilidad entre la observación superior y la inferior, pero tiene algunas aplicaciones útiles.
- En la industria, el rango se puede determinar al especificar por adelantado que una medición particular de los artículos que salen de una línea de producción deba caer dentro de cierto intervalo.

Algunos Estadísticos Importantes – Variabilidad en la Muestra (cont.)

- Si $X_1, X_2, \ldots, X_n$ representan una muestra aleatoria de tamaño n, entonces la **varianza de la muestra** se define mediante el estadístico

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

- El valor calculado de S^2 para una muestra dada se denota con s^2.
- La varianza se define, esencialmente, como el promedio de los cuadrados de las desviaciones de las observaciones de su media.

Algunos Estadísticos Importantes – Variabilidad en la Muestra (cont.)

- **Teorema.** Si S^2 es la varianza de una muestra aleatoria de tamaño n, se puede escribir como

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

Algunos Estadísticos Importantes – Variabilidad en la Muestra (cont.)

- La **desviación estándar de la muestra**, que se denota con *S*, es la raíz cuadrada positiva de la varianza de la muestra.

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

- La cantidad *n* – 1 a menudo se denomina **grados de libertad asociados con la varianza** estimada. Los grados de libertad representan el número de piezas de información independientes disponibles para calcular la variabilidad.

Presentaciones de Datos y Métodos Gráficos

- En la estadística, con frecuencia se hace la suposición de que la distribución es normal.
- La información gráfica con respecta a la validez de esta suposición se puede obtener de presentaciones como los diagramas de tronco y hojas, y los histogramas de frecuencias.
- A continuación se introduce la noción de gráficas de probabilidad normal y gráficas de cuantiles.
    - Estas gráficas se utilizan en estudios que tienen grados de complejidad que varían, con el objetivo principal de que las gráficas proporcionen una verificación diagnóstica de la suposición de que los datos vienen de una distribución normal.

Presentaciones de Datos y Métodos Gráficos (cont.)

- Los estadísticos vistas anteriormente proporcionan medidas simples, mientras que una representación gráfica agrega información adicional en términos de una imagen.
  - Las muestras múltiples se pueden comparar de forma gráfica.
  - Las gráficas de datos pueden sugerir relaciones entre variables.
  - Las gráficas pueden ayudar en la detección de anomalías o de observaciones de datos apartados en las muestras.

Presentaciones de Datos y Métodos Gráficos – Gráfico de Caja y Extensión (cont.)

- Esta gráfica encierra el **rango intercuartil** de los datos en una caja que tiene la mediana representada dentro.
- El rango intercuartil tiene como extremos el percentil 75 (cuartil superior) y el percentil 25 (cuartil inferior).
- Además, de la caja se prolongan extensiones, que muestran las **observaciones extremas** en la muestra.
- Para muestras razonablemente grandes, la presentación muestra el centro de la localización, la variabilidad y el grado de asimetría.

Presentaciones de Datos y Métodos Gráficos – Gráfico de Caja y Extensión (cont.)

- Una variación que se llama **gráfica de caja** puede proporcionar a quien la ve información con respecto a cuales observaciones **son datos apartados**.
  - Los datos apartados son observaciones que se consideran inusualmente alejadas de la masa de datos.
  - Técnicamente, se puede considerar un dato apartado como una observación que representa un “evento raro”; es decir, existe una probabilidad pequeña de obtener un valor tan alejado de la masa de datos.

Presentaciones de Datos y Métodos Gráficos – Gráfico de Caja y Extensión (cont.)

- **Ejemplo.** Los valores de nicotina de 40 cigarrillos son:

1,09	1,92	2,31	1,79	2,28
1,74	1,47	1,97	0,85	1,24
1,58	2,03	1,70	2,17	2,55
2,11	1,86	1,90	1,68	1,51
1,64	0,72	1,69	1,85	1,82
1,79	2,46	1,88	2,08	1,67
1,37	1,93	1,40	1,64	2,09
1,75	1,63	2,37	1,75	1,69

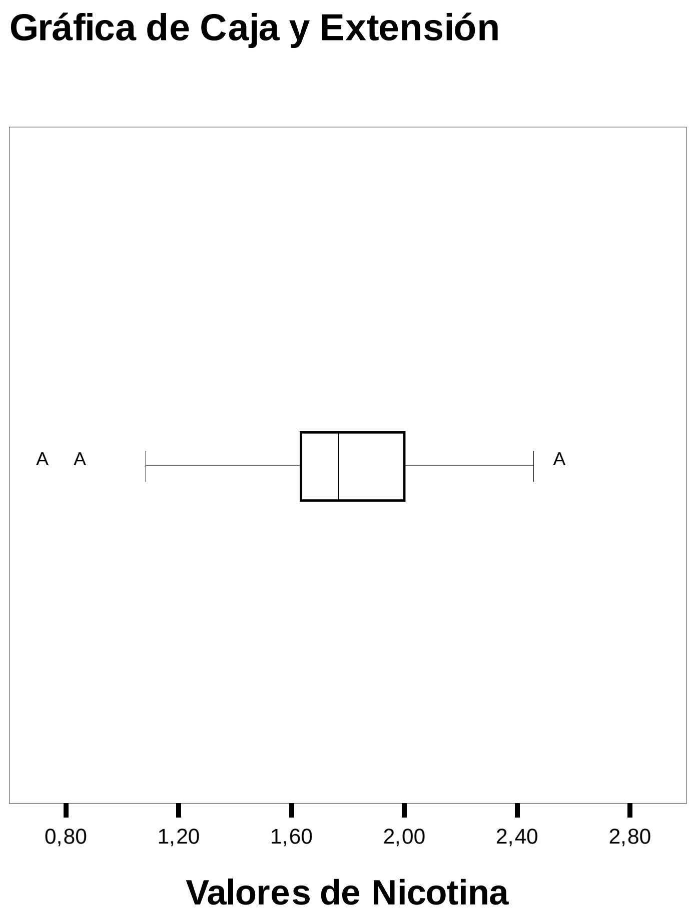
Presentaciones de Datos y Métodos Gráficos – Gráfico de Caja y Extensión (cont.)

- Se tienen las siguientes estadísticas:

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std. Deviation	Variance
Valores de Nicotina	40	1,83	,72	2,55	1,7743	,39046	,152

Presentaciones de Datos y Métodos Gráficos – Gráfico de Caja y Extensión (cont.)



Presentaciones de Datos y Métodos Gráficos – Gráfica de Cuantiles (cont.)

- El propósito de estas gráficas es describir, en forma de muestra, la función de distribución acumulada que se presentó en capítulos anteriores.
- Un **cuantil** de una muestra, $q(f)$, es un valor para el que una fracción específica f de los valores de los datos es menor que o igual a $q(f)$.
- Un cuantil representa una estimación de una característica de una población, o más bien, la distribución teórica.
- La mediana de la muestra es $q(0.5)$, el cuartil superior (percentil 75) es $q(0.75)$ y el cuartil inferior (percentil 25) es $q(0.25)$.

Presentaciones de Datos y Métodos Gráficos – Gráfica de Cuantiles (cont.)

- Una gráfica de cuantiles simplemente grafica los valores de los datos en el eje vertical contra una evaluación empírica de la fracción de observaciones excedidas por los valores de los datos.
- Para la propósitos teóricos esta fracción se calcula con

$$f_i = \frac{i - \frac{3}{8}}{n + \frac{1}{4}}$$

donde *i* es el orden de las observaciones cuando se clasifican de inferior a superior.

Presentaciones de Datos y Métodos Gráficos – Gráfica de Cuantiles (cont.)

- A diferencia de la gráfica de caja y extensión, la gráfica de cuantiles realmente muestra todas las observaciones.
- Todos los cuantiles, incluida la mediana y los cuantiles inferior y superior, se pueden aproximar de forma visual.
- Las indicaciones de agrupaciones relativamente grandes alrededor de valores específicos se indican por pendientes cercanas a cero, mientras que los datos dispersos en ciertas áreas producen pendientes más abruptas.

Presentaciones de Datos y Métodos Gráficos – Gráfica de Cuantiles-Cuantiles Normales (cont.)

- La gráfica de cuantiles-cuantiles normales toma ventaja de lo que se conoce acerca de los cuantiles de la distribución normal.
- La metodología incluye una gráfica de los cuantiles empíricos recién presentados contra el cuantil correspondiente de la distribución normal.

Presentaciones de Datos y Métodos Gráficos – Gráfica de Cuantiles-Cuantiles Normales (cont.)

- La expresión para un cuantil de una variable aleatoria $N(\mu,\sigma)$ es muy complicada. Una buena aproximación está dada por:

$$q_{\mu,\sigma}(f) = \mu + \sigma\left\{4.91\left[f^{0.14} - (1-f)^{0.14}\right]\right\}$$

- La expresión para un cuantil de una variable aleatoria $N(0,1)$ es:

$$q_{0,1}(f) = \left\{4.91\left[f^{0.14} - (1-f)^{0.14}\right]\right\}$$

Presentaciones de Datos y Métodos Gráficos – Gráfica de Cuantiles (cont.)

- La gráfica de cuantiles-cuantiles normales es una gráfica de $y_{(i)}$ (observaciones ordenadas) contra $q_{0,1}(f_i)$, donde

$$f_i = \frac{i - \frac{3}{8}}{n + \frac{1}{4}}$$

- Una relación cercana a una línea recta sugiere que los datos provienen de una distribución normal.
- La intersección en el eje vertical es una estimación de la media de la población y la pendiente es una estimación de la desviación estándar.

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

$$Y = a_1X_1 + a_2X_2 + \ldots + a_nX_n$$

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 σ^2_1 y σ^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.

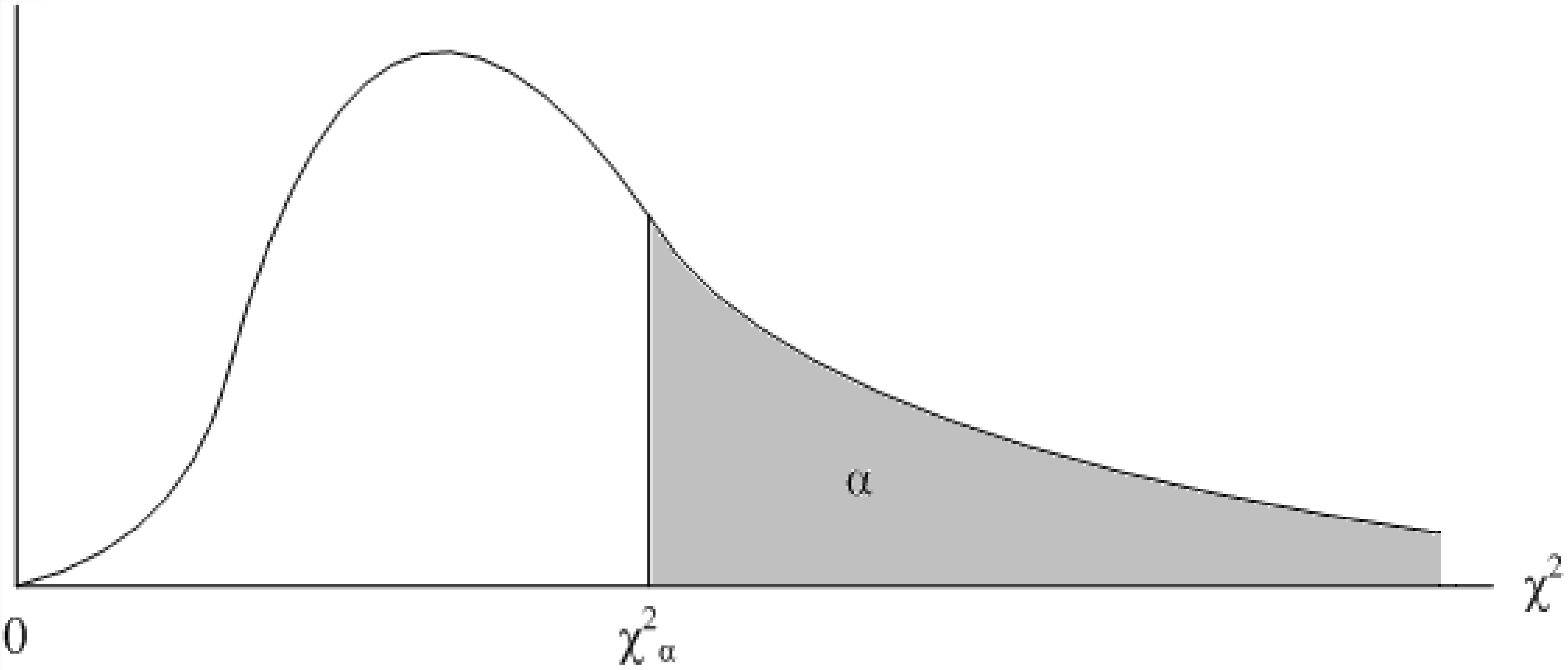
Distribución Muestral de S^2

- 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

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

  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

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

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

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

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, …, 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.

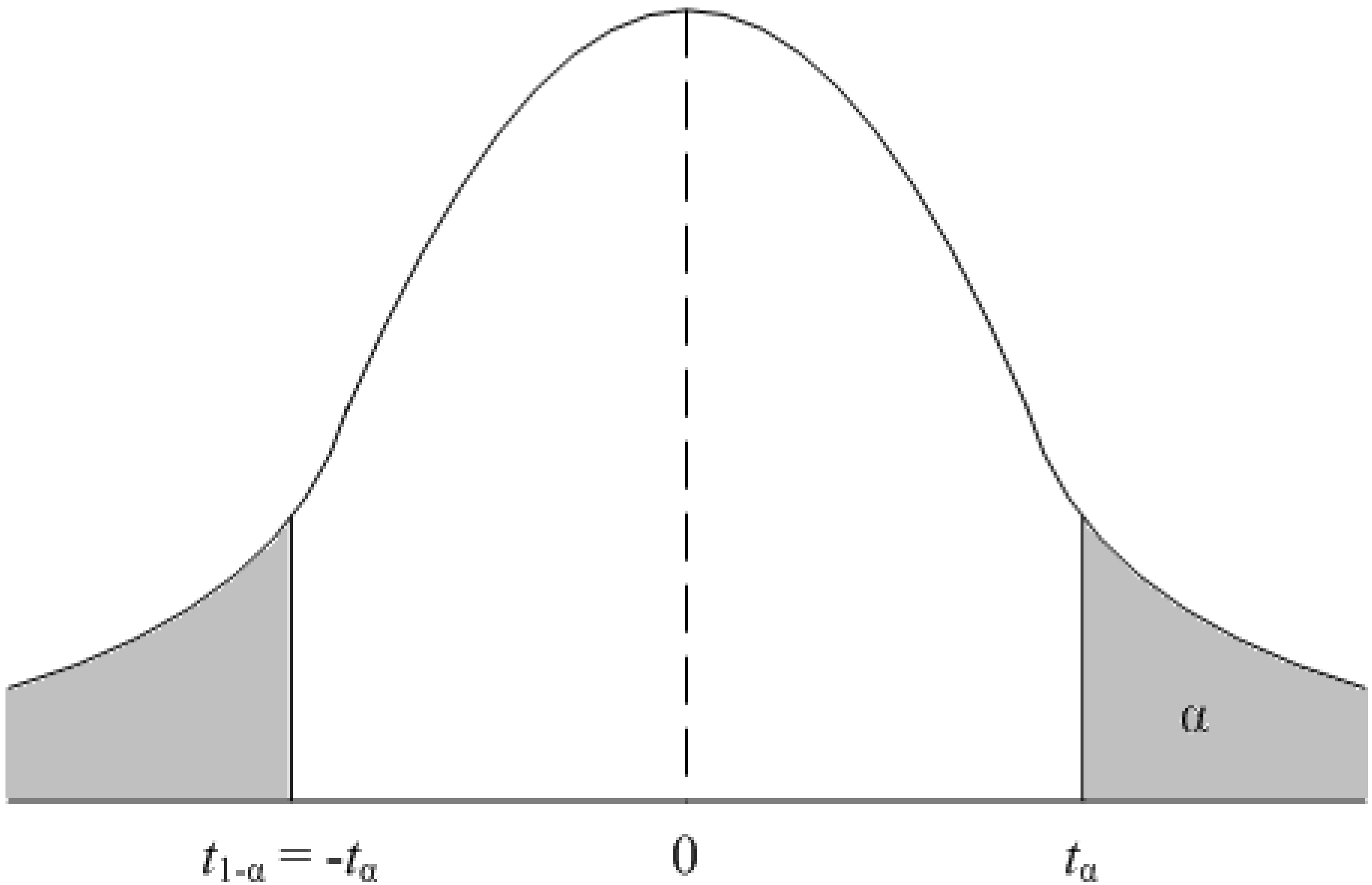
Distribución *t* (cont.)

- 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

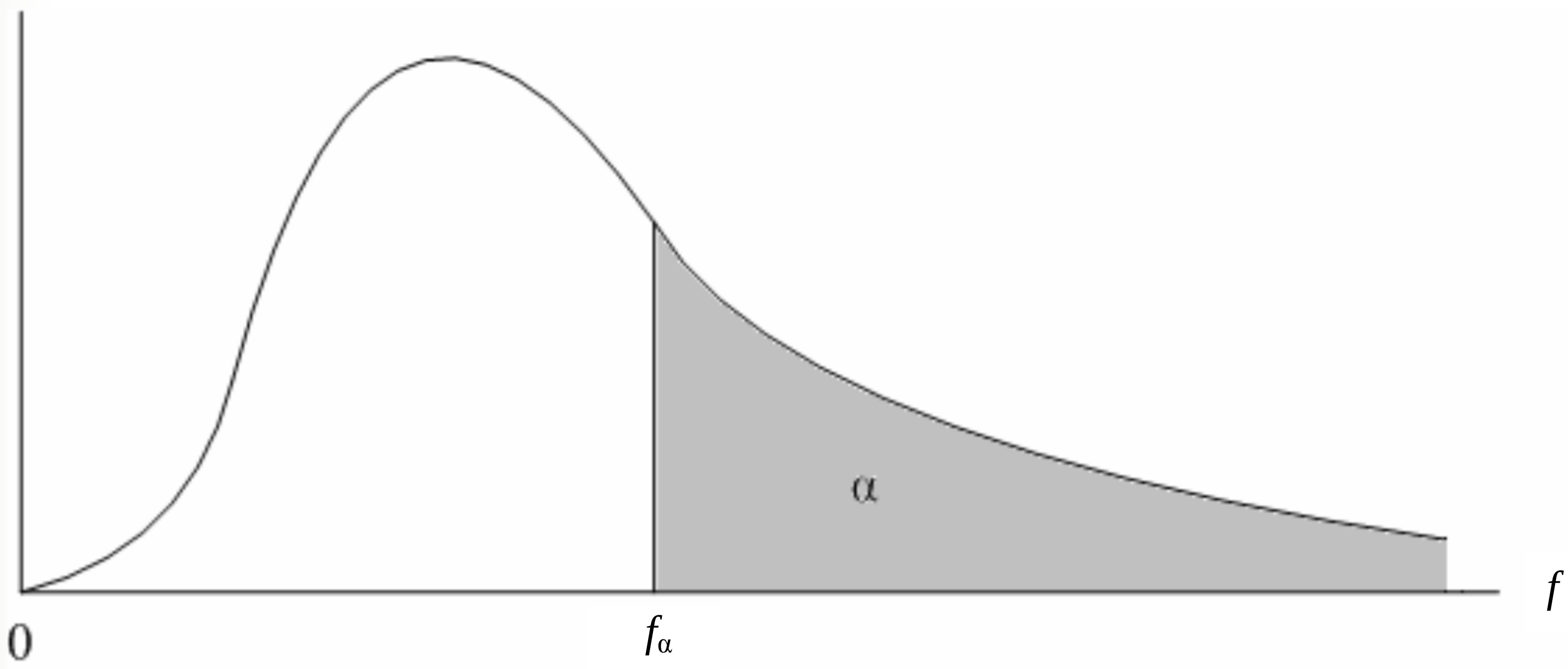
$$h(f) = \begin{cases} \dfrac{\Gamma[(v_1+v_2)/2](v_1/v_2)^{v_1/2}}{\Gamma(v_1/2)\Gamma(v_2/2)} * \dfrac{f^{v_1/2-1}}{(1+v_1 f/v_2)^{(v_1+v_2)/2}} & 0 < f < +\infty \\ 0 & \text{otro caso} \end{cases}$$

  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 $\sigma^2{}_1$ y $\sigma^2{}_2$, 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.