CAPITULO III

ESTADÍSTICA DESCRIPTIVA.

Si se mide la corriente que circula por un alambre de cobre delgado, lo que se está haciendo es un experimento. Sin embargo, al repetir la medición durante varios días, los resultados que se obtienen son un poco diferentes debido a pequeñas variaciones en las variables que no están controladas en el experimento, como son los cambios de temperatura ambiente, ligeras variaciones en el instrumento de medición y pequeñas impurezas en la composición química del alambre en distintas partes, además de las variaciones de la fuente de corriente. En consecuencia se dice que este experimento, así como muchos otros, tiene un componente **aleatorio**. En algunos casos, las variaciones aleatorias observadas son tan pequeñas en relación con las metas del experimento, que pueden ignorarse. Sin embargo, la variación casi siempre está presente y su magnitud puede llegar a ser tan importante a tal grado, que las conclusiones del experimento no sean muy evidentes

Otro ejemplo de experimento es la selección de una pieza de la producción de un día y la medición con bastante exactitud de la longitud de está. En la práctica pueden presentarse pequeñas variaciones de las longitudes de las medidas, por muchas causas, tales como vibraciones, fluctuaciones de temperatura, diferencias entre quienes toman las mediciones, calibraciones, desgastes en la herramienta de corte, desgaste en los cojinetes y cambios en la materia prima. Incluso el procedimiento de medición puede producir variaciones en los resultados finales.

En estos tipos de experimentos, las mediciones de interés, (la corriente en el alambre de cobre, la longitud de una pieza maquinada), pueden representarse con una variable aleatoria. Es razonable modelar el rango de los valores posibles de la variable aleatoria con un intervalo (finito o infinito) de números reales. Por ejemplo, para la longitud de una parte maquinada, este modelo permite que las mediciones del experimento produzcan cualquier valor dentro de un intervalo de números reales. Este intervalo puede concebirse como un continuo de valores , en consecuencia se define que “si el rango de una variable aleatoria *X* contiene un intervalo (ya sea infinito o finito) de números reales, entonces *X* es una variable aleatoria continua.

3.1 Variable aleatoria

Si arrojamos dos dados, sabemos que la suma *X* de los puntos que caen hacia arriba debe ser un número entero entre 2 y 12, pero no podemos predecir que valor de *X* aparecerá en el siguiente ensayo y podemos decir que *X* depende del azar. El tiempo de vida de un foco que se extrae aleatoriamente de un lote de focos depende también del azar.

Si las observaciones no se dan en términos de números, podemos asignarles números y reducir las observaciones cualitativas al caso cuantitativo. Por ejemplo, si se lanza una moneda 3 veces, el número de “caras” es una variable aleatoria *X* que toma los valores 0, 1, 2 ó 3 (que representan en número de veces que se obtiene “caras” en los 3 lanzamientos de la moneda). Así tenemos que la función que asigna números o valores a cada uno de los elementos del espacio muestra con una probabilidad definida, se denomina *variable aleatoria*.

El espacio de muestra es el dominio de la función y el conjunto de valores que la variable puede tomar es el rango de la función, que es un subconjunto de números reales. Si el rango de *X* es el conjunto de números enteros *Z* o un subconjunto de *Z*, la variable aleatoria se llama **variable aleatoria discreta**, y si el rango es el conjunto de números reales, *R,* o un subconjunto de *R*, la variable aleatoria se llama **variable aleatoria continua**. Son ejemplos de variables aleatorias continuas: la estatura, el peso, la edad, el volumen, el pH, etc. Algunos ejemplos de variables discretas son : el número de alumnos en una clase, el número de accidentes de automóvil, número de piezas defectuosas por lote, etc.

La posibilidad de ocurrencia de un valor para la variable aleatoria se determina en términos de su *probabilidad*. Supóngase un suceso *E*, que de un total de *n* casos posibles, todos igualmente factibles, puede presentarse en *h* de los casos. Entonces la probabilidad de aparición del suceso (llamada su ocurrencia) viene dada por:

$$p = P\{E\} = \frac{h}{n}$$

Ejemplo: Sea *E* el suceso de que aparezcan los números 3 ó 4 en una sola tirada de un dado. Hay seis casos que pueden presentarse, que son: 1, 2, 3, 4, 5 y 6. Los seis casos son igualmente posibles. Puesto que *E* puede presentarse con dos de estos casos, entonces: $p = P\{E\} = 2/6 = 1/3$

Debe tenerse muy en cuenta que la probabilidad de un suceso es un número comprendido entre 0 y 1. Si el suceso es imposible (no puede ocurrir) su probabilidad es cero. Si es un suceso cierto (tiene que ocurrir) su probabilidad es uno.

La naturaleza del estudio que se considera en el presente curso, condiciona a que solo se aborde el caso de variables aleatorias continuas, dejando de lado el tratamiento de variables aleatorias discretas.

3.2 Distribución de variables aleatorias continuas

Una función *f(x)* es una *función de densidad de probabilidad, fdp*, de la variable aleatoria continua *X*, si para cualquier intervalo de números reales [*a,b*], se tiene:

1. $f(x) \geq 0$
2. $\int_{-\infty}^{\infty} f(x)\,dx = 1$
3. $P(a \leq X \leq b) = \int_{a}^{b} f(u)\,du$

Es decir, la probabilidad $P(a \leq X < b)$ es el área sombreada de la gráfica de *f(x),* Figura 3.1, para las líneas verticales $x = a$ y $x = b$. Esta área da la probabilidad de que *X* se encuentre entre *a* y *b*. En cierto sentido, *f(x)* es el límite de la frecuencia relativa normalizada de un histograma al incrementarse el número de clases y cuando los intervalos de clase tienden a cero.

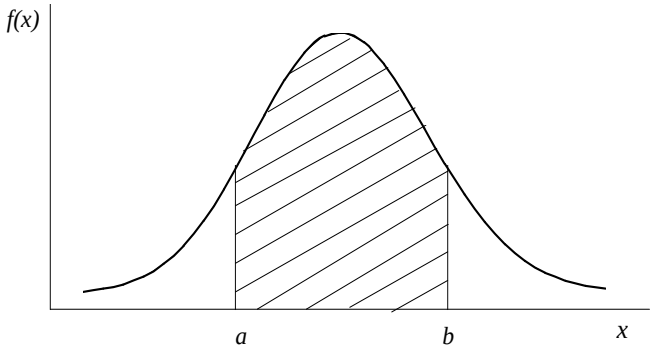


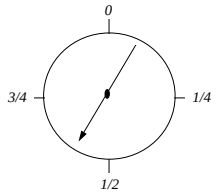
Figura 3.1: Gráfico de la densidad de probabilidad *f(x)*. El área sombreada representa $P(a \leq X < b)$.

a que áreas tales como $P(a \leq X < b)$ para toda $a < b$, representan probabilidades, se requiere que el área total debajo del gráfico de *f(x)* y *x*, sea igual a 1. Mas aún que las probabilidades son siempre positivas, se necesita que: $f(x) \geq 0;\ x \in R$. Es interesante observar que si el espacio bajo la curva corresponde a un solo valor de *x, x = b* entonces:

$P(X = b) = \int_{b}^{b} f(x)\,dx = 0$

Esto concuerda con lo intuitivo, porque si el espacio *R* es un intervalo con infinita cantidad de puntos, la probabilidad de un solo punto en particular es cero.

Ejemplo: Consideremos una rueda con una aguja giratoria balanceada:



La aguja al ser girada se detendrá en cualquier punto entre 0 y 1. Un modelo razonable para la variable aleatoria *X* es $f(x) = 1;\ x \in R = \{x;\ 0 \leq x < 1\}$, o de otra forma:

$f(x) = 1; \quad 0 \leq x < 1$

Tal *fdp* es constante en el espacio *R*. Para este modelo la probabilidad de:

$$P(\tfrac{1}{4} \le X \langle \tfrac{1}{2}) = \int_{\frac{1}{4}}^{\frac{1}{2}} 1.\, dx = 0.25;$$

es decir, la probabilidad de que al terminar de girar la aguja se detenga entre el cuadrángulo ¼ y ½ es 0.25. De otra forma, existe 25 % de probabilidades de que la aguja se detenga en el segundo cuadrángulo de esa esfera.

Hay ciertas convenciones que se usan en el contexto de las variables aleatorias continuas. Ya que en un caso continuo P(*X*=*x*) = 0, para todo $x \varepsilon R$, se tiene que:

$P(a < X < b) = P(a \le X < b) = P(a < X \le b) = P(a \le X \le b)$

Esto es, se puede incluir o excluir los signos de igualdad en estas expresiones sin cambiar la probabilidad.

3.3 La distribución normal

La distribución normal es la más importante distribución en el estudio de la estadística, debido a que son muchos los fenómenos que son ***normalmente distribuidos***. Esta distribución fue desarrollada el siglo pasado por el matemático alemán Karl F. Gauss, de modo que la distribución normal se conoce también como distribución Gaussiana.

Si X tiene una distribución normal, con promedio μ y varianza σ^2, su *fdp* es:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]} \qquad -\infty < x < \infty \qquad (1)$$

Se debe distinguir al promedio de muestra simbolizado por $\bar{x}$ del promedio de población simbolizado por μ, y de la varianza de muestra, s^2 con la varianza de población σ^2.

La *fdp* de la distribución normal se abrevia diciendo que *X* es $N(\mu, \sigma^2)$; es decir, *X* está normalmente distribuida con promedio μ, y varianza σ^2. El gráfico de *f(x)* es la bien conocida curva de campana o *curva de Gauss* mostrada en la Figura 3.2. El gráfico de $N(\mu, \sigma^2)$ es simétrico con respecto a $x = \mu$ y alcanza su máximo valor en este punto.

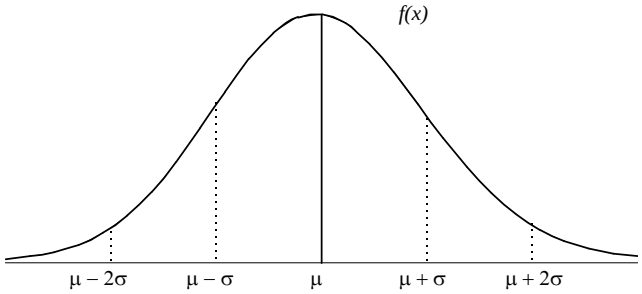


Figura 3.2: Función de densidad de probabilidad de la distribución $N(\mu, \sigma^2)$.

En general , se dice que *X* es $N(\mu, \sigma^2)$ y se quiere determinar:

$$P(a < X < b) = \int_a^b \frac{1}{\sigma\sqrt{2\pi}}\, e^{\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]} dx \qquad (2)$$

Si en la ecuación (2) hacemos que $z = (x-\mu)/\sigma$ tal que $x = \mu + \sigma z$ y $dx/dz = \sigma$, ($dx = \sigma\, dz$) se tiene que:

$$P(a < X < b) = \int_a^b \frac{1}{\sqrt{2\pi}}\, e^{\left[-\frac{z^2}{2}\right]} dz \qquad (3)$$

Se puede observar que la integral de la ecuación (3) no es fácil de determinar por lo que se recurre al uso de métodos numéricos. En Tablas aparecen tabulados los valores de esta integral para una distribución *N(0,1),* (Función Estándar de Distribución Normal) representada por:

$$\phi(z) = P(Z < z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{\left[-\frac{w^2}{2}\right]} dw$$

Una variable aleatoria normal con $\mu = 0$ y $\sigma^2 = 1$ recibe el nombre de *variable aleatoria normal estándar* y se denota como *Z*.

Las distribuciones normales sólo varían con respecto a la media y/o la desviación estándar. La media determina la posición de una curva sobre el eje horizontal. La desviación estándar determina el grado de amplitud o dispersión entre los elementos. La Figura 3.3 (a) muestra dos distribuciones normales con idénticas desviaciones estándar, pero con medias distintas. La Figura 3.3 (b) muestra dos distribuciones normales con idénticas medias y diferentes desviaciones estándar

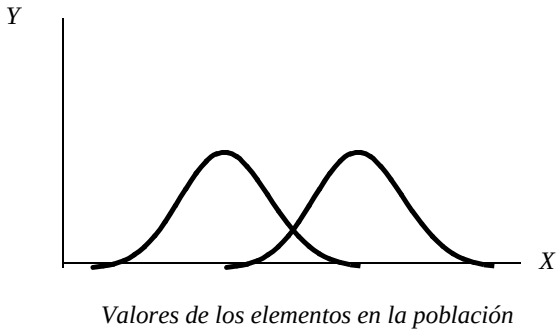


Figura 3.3 (a): Distribuciones estándar: medias diferentes y desviaciones estándar iguales.

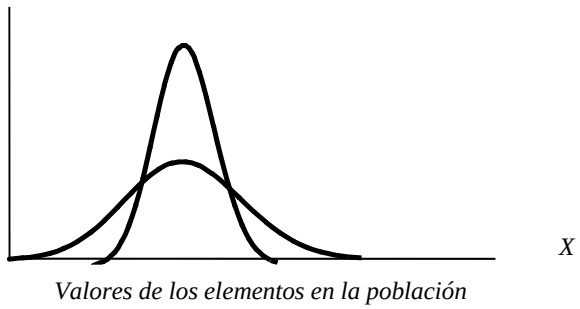


Figura 3.3 (b): Distribuciones estándar: medias iguales y desviaciones estándar distintas.

Hay un número infinito de funciones de densidad normal, una para cada combinación de μ y σ. La media μ mide la ubicación de la distribución y la desviación estándar σ mide la dispersión.

No es posible obtener una expresión de forma cerrada par la integral de la función de densidad normal. Sin embargo, se puede calcular el área debajo de la curva normal utilizando procedimientos de aproximación. Se dice entonces que:

Si *X* es una variable aleatoria normal con media μ y varianza σ^2, entonces:

$$Z = \frac{X - \mu}{\sigma}$$

es una variable aleatoria normal con media cero y varianza 1. La variable aleatoria *Z* se denomina **variable normal estándar.**

Las áreas de la variable normal estándar se dan en la Tabla A de los apéndices. Son las áreas bajo la curva normal entre $z = -\infty$ y un valor cualquiera de z, valores que definen la probabilidad de algún evento.

Por ejemplo, la probabilidad $\phi(1.5) = 0.932$ corresponde al área sombreada de la Figura 3.4

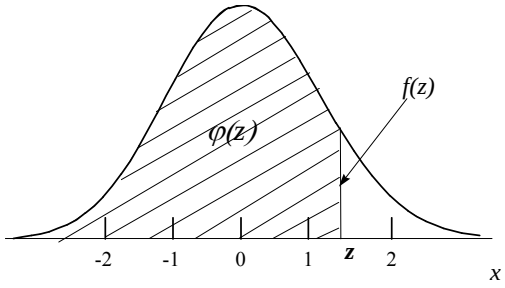


Figura 3.4: Función de densidad de probabilidad.

Con esta notación se puede escribir una probabilidad determinada, como por ejemplo:

$P(-1 < Z < 1.5) = \phi(1.5) - \phi(-1.0)$

o sea, se determina el área hasta 1.5 en la *fdp* y se resta el área de la curva de $-\infty$ a -1. Debido a la simetría de *f(x)* alrededor de *z*, es correcto que $\phi(-1.0) = 1 - \phi(1.0)$, o en términos más generales:

$\phi(-z) = 1 - \phi(z)$

Así, se puede determinar:

$P(-1 < Z < 1.5) = \phi(1.5) - [1 - \phi(1.0)]$

$= 0.9332 - (1 - 0.8413) = 0.7745$

Lo anterior corresponde a la distribución estándar $N(0,1)$. Supóngase ahora de que X es $N(\mu = 75, \sigma^2 = 100)$ y queremos determinar $P(70 < X < 90)$. En estos casos, la Tabla respectiva puede ser utilizada según:

$$P(a < Z < b) = \phi\left(\frac{b-\mu}{\sigma}\right) - \phi\left(\frac{a-\mu}{\sigma}\right)$$

Esto es, se puede estandarizar la distribución en referencia para una distribución $N(0, 1)$.

Ejemplo: si X $N(75, 100)$, entonces:

$$P(70 < X < 90) = \phi\left(\frac{90-75}{10}\right) - \phi\left(\frac{70-75}{10}\right)$$

$= \phi(1.5) - \phi(-0.5) = 0.6247$

$$P(80 < X < 95) = \phi\left(\frac{95-75}{10}\right) - \phi\left(\frac{80-75}{10}\right)$$

$= \phi(2.0) - \phi(0.5)$

$= 0.9772 - 0.6915 = 0.2857$

$$P(55 < X < 65) = \phi\left(\frac{65-75}{10}\right) - \phi\left(\frac{55-75}{10}\right)$$

$= \phi(-1.0) - \phi(-2.0)$

$= 0.9772 - 0.8413 = 0.1359$

Ejemplo: El peso promedio de mineral en un camión es de 25 toneladas, según se ha determinado de los pesos netos de mineral en 100 camiones muestreados. La desviación estándar es 5 ton. Suponiendo que la variable peso de camión esta distribuida normalmente, a) ¿Cuántos camiones contienen entre 20 y 30 toneladas de mineral, b) ¿Cuántos camiones contienen más de 40 toneladas?

Se considera que $\mu = 25$ ton y $\sigma = 5$ ton.; luego:

a) $P(20 < X < 30) = \phi\left(\frac{20-25}{5}\right) - \phi\left(\frac{30-25}{5}\right) = \phi(1) - \phi(-1) = 0{,}6826$

Entonces ~ 68 camiones contendrán entre 20 y 30 toneladas de mineral.

b) $P(X > 40) = 1 - \phi\left(\frac{40-25}{5}\right)$ $1 - \phi(3) = 0{,}0013$

Luego, 0.0013 * 100 = 0,13 ~ 0 camiones contienen más de 40 ton.

La Figura 3.5 presenta un resumen de varios resultados útiles relacionados con la distribución normal. Para cualquier variable aleatoria normal:

$P(\mu - \sigma < X < \mu + \sigma) = 0.6827$

$P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.9545$

$P(\mu - 3\sigma < X < \mu + 3\sigma) = 0.9973$

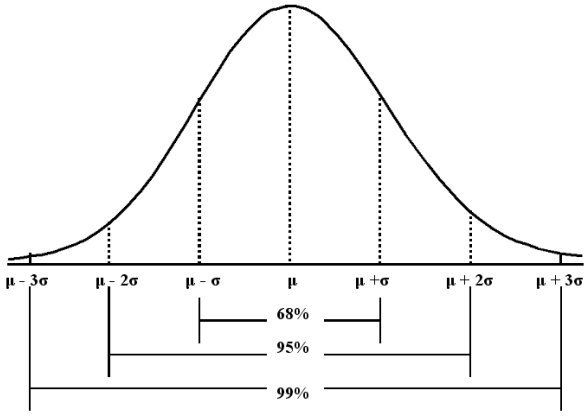


Figura 3.5: Probabilidades asociadas a una distribución normal

Debido a que más del 0.9973 de la probabilidad de una distribución normal está comprendida en el intervalo:

$(\mu - 3\sigma < X < \mu + 3\sigma)$,

a menudo se hace referencia a la cantidad 6σ como el ancho de la distribución normal. El área que se está más allá de 3σ de la media es muy pequeña

Un mejor entendimiento de la distribución normal y de sus parámetros μ y σ se logra con lo siguiente evaluación de probabilidades. Si X es $N(\mu, \sigma^2)$, para un valor $k > 0$ tenemos que:

$$P(\mu - k\sigma < X < \mu + k\sigma) = \phi\left[\frac{(\mu + k\sigma) - \mu}{\sigma}\right] - \phi\left[\frac{(\mu - k\sigma) - \mu}{\sigma}\right] = 2\phi(k) - 1$$

Para valores selectos de k, se obtiene las siguientes probabilidades:

k	1.0	1.282	1.645	1.96	2.0	2.576	3.0
$P(\mu - k\sigma < X < \mu + k\sigma)$	0.6826	0.80	0.9	0.95	0.9544	0.99	0.9974

3.4. Distribución de muestra.

Se ha estimado parámetros como el promedio μ y la desviación estándar, σ de una distribución normal, basados en observaciones $x_1, x_2, ..., x_n$ que fueron obtenidos por muestreo de una población de interés. Sin embargo hay que reconocer que generalmente estos ***estimados*** no son iguales a los verdaderos valores de la población considerada. Es decir: $\bar{x} \neq \mu$; $s \neq \sigma$. De esto resulta por ejemplo, que si se repite varias veces el muestreo de una misma población y de cada muestreo de obtiene $\bar{x}$ y s, cada uno de los respectivos valores diferirán entre si. Si realizamos N muestreos, se obtendrá $\bar{x}_1, \bar{x}_2, ..., \bar{x}_n$ promedios. Estos estimadores ($\bar{x}$, s^2, s,) por lo tanto tendrán una distribución, de lo que resulta que es necesario evaluar la confiabilidad de los estimadores. Se estará hablando por ejemplo de la varianza de promedios. Si esta varianza es muy grande, no se tendrá mucha confianza en la evaluación hecha mediante varios muestreos. Todo esto da sentido a la expresión distribución de muestra o distribución de estimadores de muestreos.

El muestreo introduce variabilidad en los estimadores. Esta fuente de variabilidad se denomina *variabilidad de muestreo* o *variabilidad debido al muestreo*.

3.4.1 Distribución del promedio de muestra, $\bar{x}$

Considerando el promedio de una muestra de tamaño n:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_1 = \left(\frac{1}{n}\right) X_1 + \left(\frac{1}{n}\right) X_2 + \ldots + \left(\frac{1}{n}\right) X_n$$

tomada de una población de media μ y varianza σ^2, entonces $\bar{X}$ es un valor de una variable aleatoria cuya distribución tiene media μ. Para muestras de población infinitas, la varianza de esta distribución es σ^2/n, o lo que es lo mismo:

$$E(\bar{X}) = \mu; \quad VAR(\bar{X}) = \frac{\sigma^2}{n}; \quad D.E.(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

esto es, el promedio de muestra $\bar{x}$ es el mismo que el de la distribución. Sin embargo, la varianza es la misma de la distribución, pero dividida por el tamaño de la muestra. El promedio de muestra $\bar{x}$ es el más común estimador del promedio de población μ, ambos valores, ($\bar{x}$ y μ) diferirán entre si cada vez que se evalúe un promedio de muestra.

El hecho de que VAR($\bar{x}$) = σ^2/n muestra que la variabilidad del estimador $\bar{x}$ alrededor del promedio μ tiende a cero según que el número de observaciones en la muestra crezcan. Hacia el límite, cuanto más grande sea *n*, el promedio de la población μ quedará determinado con mayor precisión.

Ahora bien, si $\bar{X}$ es un valor de una variable aleatoria de tamaño *n*, cuya distribución tiene media μ y varianza σ^2 entonces:

$$Z = \left(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \right)$$

es el valor de una variable aleatoria cuya función de distribución se aproxima a la de la distribución normal centrada y estandarizada (tipificada); es decir, $\bar{X}$ será *N(0, 1)*. Esto implica que la combinación lineal de variables aleatorias es también normalmente distribuida. Por lo tanto, si se toma una muestra de una distribución normal con promedio μ y varianza σ^2 entonces la distribución de $\bar{X}$ es:

$$N\left(\mu, \frac{\sigma^2}{n}\right)$$

y la distribución de *Z* es *N*(0, 1)

3.4.2 Intervalos de confianza para Promedios

Se trata de determinar un intervalo en el que se encuentre con *alta probabilidad* el promedio desconocido μ. Se utiliza la distribución de $\bar{X}$ que es aproximadamente N(μ, σ^2 /n) para valores grandes de *n*.

Para una determinada probabilidad, (1-α) se puede encontrar un valor tal como z (α/2) de la tabla normal, tal que:

$$P\left[-z(\alpha/2) \le \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \le z(\alpha/2) \right] = 1 - \alpha$$

ya que *Z* es una distribución simétrica, *N*(0, 1) alrededor de cero, entonces:

$$P\left[Z \le -z(\alpha/2)\right] = P\left[Z \ge z(\alpha/2)\right] = \alpha/2$$

Generalmente α es una valor pequeño tal como 0.1, 0.05 ó 0.01. De este modo, si:

(1 - α) = 0.95, entonces z (α/2) = z(0.025) = 1.96; y si

(1 - α) = 0.90, entonces z (α/2) = z(0.05) = 1.645.

Las desigualdades:

$$-z(\alpha/2) \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z(\alpha/2)$$

son equivalentes a:

$$\bar{X} - z(\alpha/2)\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z(\alpha/2)\frac{\sigma}{\sqrt{n}}$$

Así, las probabilidades para cada una de estas desigualdades es (1 - α). En particular:

$$P\left[\bar{X} - z(\alpha/2)\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z(\alpha/2)\frac{\sigma}{\sqrt{n}}\right] = 1-\alpha$$

Obsérvese que $\bar{X}$ esta en los extremos de las desigualdades y el parámetro constante, pero desconocido esta en el medio. Así, la probabilidad de que el intervalo aleatorio;

$$\left[\bar{X} - z(\alpha/2)\frac{\sigma}{\sqrt{n}},\ \bar{X} + z(\alpha/2)\frac{\sigma}{\sqrt{n}}\right]$$

incluya el promedio desconocido μ, es (1 - α). Para simplificar el intervalo, se le puede escribir:

$$\bar{X} \pm z(\alpha/2)\frac{\sigma}{\sqrt{n}}$$

3.4.3 Intervalos de confianza para μ_1 - μ_2

Sean los promedios de muestra $\bar{X}$ y $\bar{Y}$ estimadores de μ_1 y μ_2 con varianzas σ_1^2 y σ_2^2 respectivamente. Si las muestras son tomadas de distribuciones normales, las distribuciones respectivas para esos promedios serán:

$$N(\mu_1, \sigma_1^2/n_1) \quad y \quad N(\mu_2, \sigma_2^2/n_2)$$

El estimador apropiado para μ_1 - μ_2 es la diferencia de los promedios de muestra $\bar{X}$ y $\bar{Y}$. Se debe asumir que se trata de dos muestras aleatorias seleccionadas ***independientemente***. Por consiguiente:

$$Var(\bar{X}-\bar{Y}) = Var(\bar{X}) + Var(\bar{Y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

por lo que la distribución de muestra de $\bar{X} - \bar{Y}$ es:

$$N\left(\mu_1 - \mu_2,\ \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

con lo que se puede deducir que:

$$\left[\bar{X} - \bar{Y} - z(\alpha/2)\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},\ \bar{X} - \bar{Y} + z(\alpha/2)\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right]$$

incluye μ_1 - μ_2 a 100(1-α) por ciento de confianza.

Si las varianzas de muestra (s^2) son conocidas y las varianzas de población (σ^2) desconocidas, es posible que el intervalo:

$$\left[\bar{x}-\bar{y}\pm z(\alpha/2)\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}\right]$$

proporcione aproximadamente al 100(1-α) por ciento de confianza para $\mu_1 - \mu_2$

3.5. Inferencias con pequeñas muestras y varianzas desconocidas

El uso de la teoría anterior requiere que se conozca σ. Si *n* es grande, se puede usar también esa teoría cuando no se conoce σ y puede ser reemplazada por *s.* Para muestras de tamaño ***n* < 3 0**, llamadas muestras pequeñas, esta aproximación no es muy buena y va siendo tanto peor a medida que *n* disminuya; por lo tanto, no se puede tener mucha confianza en *"s"* como aproximación de σ.

Cuando ocurre así, se puede probar el siguiente teorema:

"Si $\bar{X}$ es la media de una muestra aleatoria de tamaño *n* tomada de una población normal con la media μ y varianza σ^2, entonces:

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

es el valor de una variable aleatoria que tiene una distribución *t-Student* de parámetro *r = n - 1 grados de libertad*.

En este caso no se requiere conocer σ y se debe trabajar con una población normal.

La forma general de una distribución "*t"* es similar a la de una distribución normal; ambas tienen la forma de campana y son simétricas con respecto a la media; el valor máximo de l a ordenada se alcanza en la media μ = 0. Sin embargo, la distribución *t* tiene colas más amplias que la normal; esto es, la probabilidad de las colas es mayor que en la distribución normal. A medida que el número de grados de libertad, $r \to \infty$ la forma límite de la distribución t es la distribución normal; Figura 3.6.

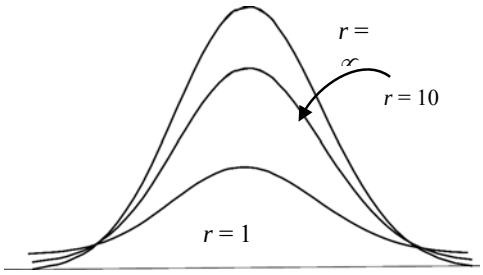


Figura 3.6: Funciones de densidad de probabilidad para varias distribuciones *t.*

Como la distribución normal tipificada o estandarizada, la distribución *t* tiene media cero, pero su varianza depende del parámetro *r* (υ según nomenclatura de algunos autores) llamado número de grados de libertad. La *fdp* para la distribución *t-student* con *r* grados de libertad es:

$$f(t)=\frac{c}{\left(1+t^2/r\right)^{(r+1)/2}} \qquad -\infty < t < \infty$$

donde *c* es un valor tal que el área debajo de *f(t)* = 1. Se comprueba que *E(t)* = 0 y *Var(t)* = *r/r*-2 para r>2. La varianza de *t* es mayor de 1 pero se aproxima a ese valor cuando $n \to \infty$. Esta densidad se parece mucho a la distribución *N*(0, 1), especialmente para valores grandes de *r*.

En la Tabla respectiva figuran los porcentajes mayores de probabilidades de cola, tal que:

$P[T > t(\alpha; r)] = \alpha$

Ya que la distribución de *f(t)* es simétrica alrededor de cero, se tiene:

$P[T < -t(\alpha; r)] = P[T > t(\alpha; r)]$

3.5.1 Intervalos de confianza para promedios

Con el hecho de que:

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

t(r) = *t(n*-1), se puede escribir:

$$P\left[-t(\alpha/2; n-1) \le \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \le t(\alpha/2; n-1) \right] = 1-\alpha$$

con lo que se tiene:

$$P\left[\bar{x} - t(\alpha/2; n-1)\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t(\alpha/2; n-1)\frac{s}{\sqrt{n}} \right] = 1-\alpha$$

o sea que el intervalo;

$$\bar{x} \pm t(\alpha/2; n-1)\frac{s}{\sqrt{n}}$$

proporciona un 100(1-α) de confianza para el intervalo en que se encuentre μ. De otro modo, ese es el intervalo con (1-α) de probabilidad para encontrar μ.

3.5.2 Intervalos de confianza para $\mu_1 - \mu_2$

Por consideraciones semejantes que las señaladas para grandes muestras, se puede demostrar que el intervalo que provee del 100((1-α) por ciento de confianza para el intervalo de $\mu_1 - \mu_2$, tratándose ahora de pequeñas muestras es:

$$\bar{x} - \bar{y} \pm t(\alpha/2; r = n_1 + n_2 - 2) S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Siendo:

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

3.6. Distribución χ^2, (Chi Cuadrado)

Al igual que el promedio es una variable aleatoria, la varianza es también una variable aleatoria con una distribución muestral. La distribución muestral teórica de s^2 se encuentra ligada a una distribución gamma de parámetros $\alpha = r/2$ y $\beta = 2$ llamada distribución Chi Cuadrado (χ^2). Como s^2 no puede ser negativa, es de esperar una distribución muestral que no sea normal. Concretando se tiene:

"Si s^2 es la varianza de una muestra aleatoria de tamaño *n* tomada de una población ***normal***, que tiene varianza σ^2, entonces:

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

es el valor de una variable aleatoria que tiene distribución χ^2 con parámetro *r* = *n* - 1 llamado grados de libertad"

En la Tabla respectiva se anotan valores seleccionados de $\chi^2(\alpha; r)$, donde el área bajo la curva de la distribución χ^2 (tomada a la derecha) es igual a α.

3.6.1 Intervalos de Confianza para la varianza

Se trata de estimar dentro de unos determinados límites de confianza la varianza y desviación estándar de la población, σ, a partir de la desviación estándar muestral, *s*.

Si s^2 es la varianza muestral de una muestra aleatoria de *n* observaciones tomadas de una distribución normal con varianza desconocida σ^2 , entonces un intervalo de confianza 100(1-α) por ciento para σ^2 es:

$$\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}$$

Por ejemplo, si $\chi^2_{0.025}$ *y* $\chi^2_{0.975}$ son los valores de χ^2 (llamados valores críticos), para que el 2.5% del área se encuentre en cada *cola* de la distribución, entonces el intervalo de confianza al 95% para la varianza es:

$$\frac{s^2(n-1)}{\chi^2_{0.975,n-1}} < \sigma^2 < \frac{s^2(n-1)}{\chi^2_{0.025,n-1}}$$

y para la desviación estándar:

$$\frac{s\sqrt{n-1}}{\chi_{0.975,n-1}} < \sigma < \frac{s\sqrt{n-1}}{\chi_{0.025,n-1}}$$

3.7 Razón de dos varianzas de muestras

Supóngase ahora que se tiene dos distribuciones normales independientes y se quiere comparar sus varianzas σ_1^2 *y* σ_2^2 definiendo un intervalo de confianza para la relación:

σ_1^2 / σ_2^2

Esto es importante cuando se quiere determinar si dos muestras tienen varianzas iguales. Si esto ocurre, las dos muestras tendrán aproximadamente la misma varianza, lo cual significa que su razón será aproximadamente 1. Supóngase que se tienen dos poblaciones normales con varianzas σ_1^2 *y* σ_2^2 respectivamente. Se toman muestras aleatorias independientes de tamaños n_1 y n_2 de las poblaiciones 1 y 2 respectivamente, y sean S_1^2 *y* S_2^2 las varianza muestrales; entonces el cociente:

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

tiene una distribución *F* con es un valor de una variable aleatoria que tiene la distribución *F* con $r_1 = n_1 - 1$ grados de libertad en el numerador y r= $n_2 - 1$ grados de libertad en el denominador.

La función de densidad de probabilidad se asemeja bastante a la χ^2 . Los porcentajes superiores de $F(\alpha; r_1, r_2)$ son tal que $P[F > F(\alpha; r_1, r_2)] = \alpha$ para probabilidades de cola $\alpha = 0.05$ y $\alpha = 0.01$ y se dan en la Tabla respectiva.

Si S_1^2 *y* S_2^2 son las varianzas muestrales de dos muestras aleatorias de tamaños n_1 y n_2 respectivamente, tomadas de dos poblaciones normales e independientes son varianzas σ_1^2 *y* σ_2^2 desconocidas entonces un intervalo de confianza 100(1-α) por ciento para σ_1^2 / σ_2^2 es:

$$\frac{S_1^2}{S_2^2} f_{1-\alpha/2;n_2-1,n_1-1} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{S_1^2}{S_2^2} f_{\alpha/2;n_2-1,n_1-1}$$

donde

$f_{1-\alpha/2;n_2-1,n_1-1}$ *y* $f_{\alpha/2;n_2-1,n_1-1}$ son los puntos críticos superior e inferior que corresponden al porcentaje $\alpha/2$ de la distribución *F* con n_2*-1 y* n_1*-1* grados de libertad en el numerador y en el denominador respectivamente.

EJERCICIOS GRUPO 3

1. Hallar el área bajo la curva de distribución normal en cada uno de los siguientes casos:
   a) Entre z = 0 y z = 1.2
   b) Entre z = -0.68 y z = 0
   c) Entre z = - 0.46 y z = 2.21
   d) A la derecha de z = 2.05 y a la izquierda de z = - 1.44
   e) A la izquierda de z = -0.6
   f) A la derecha de z = - 1.28
   g) Entre z = 0.81 y z = 1.94

   Rpta: a) 0.384; b) 0.2517; c) 0.6636

2. Sea Z N(0,1). Determine:
   a) P(-0.7 < Z < 1.3)
   b) P(0.2 < Z < 1.1)
   c) P(-1.9 < Z < -0.6)

   Rpta: a) 0.6612; b) 0.285; c) 0.2456

3. Sea X N(25,16). Determine:
   a) P(19 < X < 24)
   b) P(X >23)
   c) P(X < 25)

   Rpta: a) 0.3345; b) 0.6915; c) 0.5

4. Si X es N(5,4), encontrar el valor de *c* tal que:
   a) P(X < *c*) = 0.8749
   b) P(*c* <X) = 0.6406
   c) P(X < *c*) = 0.95

   Rpta: a) 1.15; b) –0.36; c) 1.645

5. Se tiene una población infinita con media μ = 53 y varianza σ^2 = 400. Cuál es la probabilidad de obtener una muestra entre 5 y 56?
   Rpta.: 0.5514.

6. Un fabricante de cereales en paquetes tiene una etiqueta indicando peso de 12 onzas de cereal por bolsa. Su distribución de pesos es N(12.2, 0.04). Qué porcentaje de las bolsas tienen peso de cereal inferior a las 12 onzas?
   Rpta.: 15.87 %

7. Supóngase que las especificaciones del diámetro de un eje de motor son 0.25 $\pm$ 0.002 pulgadas. Si la producción de estos ejes esta distribuida normalmente con μ = 0.251 pulg. y σ = 0.001 pulg. Qué porcentaje de los ejes se encuentran dentro de las especificaciones?
   Rpta.: 84 %

8. Si los diámetros de cojinetes de bolas se distribuyen normalmente con media 0.6140 pulgadas y desviación estándar 0.0025 pulgadas, determinar el porcentaje de cojinetes de bolas con diámetros:
   a) Entre 0.610 y 0.618 pulgadas;
   b) Mayor que 0.617 pulgadas;
   c) Menor que 0.608 pulgadas.
   Rpta.: a) 89.04 %; b) 11.51 %; c) 0.82 %

9. Una máquina llena bolsas con cemento a 100 libras. El peso que se pone en cada bolsa es una variable aleatoria con σ = 0.5 lbs. El promedio de la distribución puede ser fijado por el operador. A que peso promedio debe ser fijada la máquina para que solo 5 % de las bolsas estén bajo el peso especificado?
   Rpta.: 100.823

10. El peso promedio de cierta marca de carretillas es 31 Kg. Debido a la variabilidad de la materia prima y de las condiciones de producción, el peso de estas carretillas es una variable aleatoria. Si la distribución es normal con σ = 0.5 Kg:
    a) Cuál es la probabilidad de que una muestra de carretilla aleatoriamente seleccionada pese más de 32.0 Kg?

b) Cuál es la probabilidad de que una muestra de carretilla aleatoriamente seleccionada pese entre 30.0 y 30.5 Kg?
Rpta.: a) 2.28 %; b) 13.59 %

11. La media de los diámetros interiores de una muestra de 200 arandelas producida por una máquina es de 0.502 pulgadas y la desviación estándar 0.005 pulgadas. El propósito para el que se destinan estas arandelas permite una tolerancia máxima de en el diámetro de 0.496 a 0.508 pulgadas, de otro modo las arandelas tienen que desecharse. Determinar el porcentaje de arandelas de desecho producidas por la máquina, suponiendo que los diámetros se distribuyen normalmente.
Rpta.: 23.02 %

12. El espesor de placas metálicas es una variable de interés. Debido a muchos factores, tales como variaciones en las características del metal, diferentes operarios y diferentes máquinas, el espesor varía y puede ser considerado como una variable aleatoria con μ = 20 mm y σ = 0.04 mm. Cuánto de placas de desecho se puede esperar si el espesor :
a) Tiene que ser por lo menos 19.95 mm?
b) Puede ser máximo 20.10 mm?
c) Pueden diferir máximo 0.05 mm del objetivo de 20 mm?
d) Como se establecerían los límites de tolerancia, (20 - c) y (20 + c), de tal modo que se produzca un máximo de 5 % de desechos?
e) Asúmase que el promedio se ha desplazado a μ = 20.10 mm. Calcular el porcentaje de placas metálicas que excedan los límites de tolerancia de la parte (d) de este problema.
Rpta.: a) 10.56 %; b) 0.62 %; c) 21.12 %; d) [19.92; 20.784]; e) 70.54 %

13 En una fábrica se llenan bolsas con 12 Kg. de harina. Supóngase que la distribución del peso es *N*(12.2, 0.04). Si $\bar{X}$ es el promedio de peso de 4 bolsas seleccionadas al azar, evaluar P($\bar{X}$ < 12).
Rpta.: 0.0228

14. Una muestra aleatoria de 36 muestras de *N(μ,* 25) tiene $\bar{x}$ = 49.2. Encontrar un intervalo a 90% de confianza para *μ.*
Rpta.: [47.83; 50.57]

15 El promedio y la desviación estándar de 42 exámenes de ingreso son $\bar{x}$ = 680, s = 35. Encontrar a 99% de confianza el intervalo para el promedio poblacional.
Rpta.: [666.1; 693.9]

16. Las medidas de los diámetros de una muestra al azar de 200 cojinetes de bolas hechos por una determinada máquina durante una semana dieron un promedio de 0.824 pulgadas y una desviación estándar de 0.042 pulgadas. Hallar los límites de confianza para el diámetro medio de todos los cojinetes al:
a) 95 %; b) 99 %
Rpta.: a) [0.818; 0.8298]; b) [0.816; 0.832]

17. Una compañía tiene 500 cables. Un ensayo con 40 cables elegidos al azar dieron una media de resistencia a la rotura de 2400 lbs y una desviación estándar de 150 lb. Cuáles son los límites de confianza al 95 y 99 % para estimar la media de la resistencia a la rotura de los 460 cables restantes?
Rpta.: a) [2353.51; 2446.49]; b) [2338.81; 2461.19]

18. El promedio *μ* de resistencia a la rotura de cierto papel es de interés. De 22 muestras tomadas aleatoriamente se encontró $\bar{x}$ = 2.4 libras.
a) Si la desviación estándar de una observación individual se sabe que es σ = 0.2. Encontrar al 95 % de confianza el intervalo de confianza para *μ* .
b) Si la desviación estándar σ es desconocida pero la desviación de muestra es s = 0.2, determine al 95 % de confianza el intervalo para *μ.*
Rpta.: a) [2.3164; 2.4835]; b) [2.3113; 2.4887]

19. Si una variable *U* tiene una distribución de *t-Student* con *r* = 10, hallar el valor de la constante *C*, tal que:
a) *P(U>C)* = 0.05 b) P(-C ≤ U ≤ C) = 0.98 c) P(U ≥ C) = 0.9
Rpta.: a) 1.812; b) 2.764; c) 1.37

20. Una muestra de 12 medidas de resistencia a la rotura de hebras de algodón dio una media de 7.38 onzas y una desviación estándar de 1.24 onzas. Hallar los intervalos de confianza para la resistencia real, al:
a) 95 %; b) 99 %.

Rpta.: a) [6.59; 8.17]; b) [6.268; 8.492]

21. Cinco medidas del tiempo de una reacción química fueron registradas como 0.28, 0.30, 0.27, 0.33, 0.31 segundos. Hallar los límites de confianza para el tiempo real de reacción al:
    a) 95 %; b) 99 %.
    Rpta.: a) [0.268; 0.328]; b) [0.249; 0.347]

22. Se toma 25 muestras de una población normal con media 100 y desviación estándar 10 ¿cuál es la probabilidad de que la media muestral se encuentre en el intervalo $\mu - 1.8\sigma$ y $\mu + 1.0\sigma$.
    Rpta.: 1.0

23. En la fabricación de una alfombra se utiliza fibra sintética con una resistencia a la tensión que tiene una distribución normal con media 75.5 psi y desviación estándar 3.5 psi .
    a) Encuentre la probabilidad de que una muestra aleatoria de n=6 espécimenes de fibra, la media de la resistencia a la tensión en la muestra sea mayor que 75.75 psi.
    b) ¿Cómo cambia la desviación estándar de la media muestral cuando el tamaño de la muestra aumenta desde n=6 hasta n=49?
    Rpta.: a) 0.43 b) 0.5

24. La elasticidad de un polímero es afectada por la concentración de un reactivo. Cuando se utiliza una concentración baja, la elasticidad promedio verdadera es 55, mientras que cuando se emplea una concentración alta, la elasticidad promedio es 60. La desviación estándar de la elasticidad es 4, sin importar cual sea la concentración. Si se toman dos muestras aleatorias de tamaño 16, indique si hay efecto significativo en la elasticidad del polímero por el uso del reactivo en esas concentraciones.
    Rpta.: [-2.23; -7.77]; Si hay efecto significativo.

25. Se utilizan dos máquinas para llenar botellas de plástico con detergente para máquinas lavaplatos. Se sabe que las desviaciones estándar del volumen de llenado son 0.1 onzas y 0.15 onzas para las máquinas 1 y 2 respectivamente. Se toman dos muestras aleatorias n_1 = 12 botellas y n_2= 10 botellas, con volúmenes promedio de 30.87 y 30.68 onzas respectivamente.
    a) Construya los intervalos de confianza bilateral al 95 y 99% para la diferencia entre las medias del volumen de llenado.
    b) ¿Se puede afirmar que las dos máquinas tienen diferencia entre si en el llenado del detergente?
    Rpta.: a) [0.2998; 0.0812]; b) [0.3332; 0.047] Si existe diferencia en el llenado.

26 En un proceso químico se fabrica cierto polímero. Normalmente, se hacen mediciones de viscosidad después de cada corrida, y la experiencia acumulada indican que la variabilidad del proceso es muy estable con σ= 20. Las siguientes son 15 mediciones de viscosidad por corrida:
724; 718; 776; 760; 745; 759; 795; 756; 742; 740; 761; 749; 739; 747; 742.
Encuentre un intervalo de confianza bilateral al 90% para la viscosidad media del polímero.
Rpta.: a) 750.2 ± 8.495

27 Se piensa que la concentración del ingrediente activo de un detergente líquido para ropa es afectada por el tipo de catalizador utilizado en el proceso de fabricación. Se sabe que la desviación estándar de la concentración activa es de 3 g/l, sin importar el tipo de catalizador utilizado. Se realizan 10 observaciones con cada catalizador, y se obtienen los datos siguientes:
Catalizador 1: 57.9; 66.2; 65.4; 65.4; 65.2; 65.6; 67.6; 63.7; 67.2; 71.0
Catalizador 2: 66.4; 71.7; 70.3; 69.3; 64.8; 69.6; 68.6; 64.9; 65.3; 68.8
¿Existe alguna evidencia que indique que las concentraciones activas medias dependen del catalizador utilizado?
Rpta.: a) [-5.08; 0.1797]; No existe evidencia.

28. El administrador de un lote de automóviles prueba dos marcas de llantas radiales. Para ello asigna al azar una llanta de cada marca a las dos ruedas posteriores de ocho automóviles y luego corre los automóviles hasta que las llantas se desgastan. Los datos obtenidos en kilómetros son :

Automóvil	Marca 1	Marca 2
1	36925	34318
2	45300	42280
3	36240	35500
4	32100	31950
5	37210	38015
6	48360	47800

7	38200	37810
8	33500	32215

¿Qué llanta cree que preferirá el administrador?
Rpta.: [-3164.27; 5151.07]; Ambas rinden igual.

29. Hallar χ_1^2 y χ_2^2 tales que el área bajo la distribución $\chi^2(\alpha, r)$ correspondiente a *r* = 20 entre χ_1^2 y χ_2^2 sea 0.95 suponiendo iguales las áreas a la izquierda de χ_1^2 y a la derecha de χ_2^2
Rpta.: 34.17; 9.591

30. Si la variable *U* se distribuye en $\chi^2(\alpha, r)$ con *r* = 7, hallar χ_1^2 y χ_2^2 tales que:
a) P(*U* > χ_2^2) = 0.25;
b) P($\chi_1^2 \leq U \leq \chi_2^2$) = 0.90
c) P(*U* < χ_1^2) = 0.95
Rpta.: a) 16.013; b) 14.067; 2.167; c) 14.067.

31. La desviación estándar de la duración de 10 bombillas fabricadas por una compañía es de 120 horas. Hallar los límites de confianza para la desviación estándar de todas las bombillas fabricadas por la compañía al:
a) 95 %; b) 99 %.
Resolver este problema si para 25 bombillas se encuentra *s* = 120 horas.
Rpta.: a) [82.54; 219]; [93.7; 166.38]

32. La efectividad de dos métodos de enseñanza son comparados. Una clase de 24 estudiantes es dividida aleatoriamente en dos grupos y cada grupo recibe enseñanza según cada uno de los métodos. Al finalizar el semestre se observó lo siguiente:

n_1 = 13; $\bar{x}$ = 74.5; $s_1^2 = 82.6$

n_2 = 11; $\bar{x}$ = 71.8; $s_2^2 = 112.6$

Asumir una distribución normal con $\sigma_1^2 = \sigma_2^2$. Calcular al 90 % de confianza el intervalo para σ_1^2 / σ_2^2.
Hay diferencia significativa entre los dos métodos de enseñanza?.
Rpta.: [.01848; 2.02]; No hay diferencia.

33. Dos compuestos de caucho fueron probados para resistencia a la tracción. Se prepararon 14 muestras rectangulares, 7 para cada una de las muestras *A* y *B*. Durante la experimentación se observó que dos espécimenes de *B* estuvieron defectuosos por lo que se les eliminó de la prueba. La fuerza de tracción (en unidades de 100 psi) fueron como sigue:
A = 32 30 33 32 29 34 32
B = 33 35 36 37 35
Calcular al 90 % de confianza el intervalo para σ_1^2 / σ_2^2. Comente los resultados.
Rpta: [0.2143; 5.82]; no hay diferencia entre las dos varianzas.