1.3. Hacer estimaciones sobre una población a partir de una muestra:

Objetivos:

1. Estudiar las propiedades de las distribuciones de las muestras de una población.
2. Identificar los estadísticos-parámetros de una muestra que mejor permiten estimar los parámetros de la población
3. Construir intervalos que contengan con una cierta probabilidad el valor de un parámetro poblacional.
4. Determinar la probabilidad de equivocarnos al rechazar una hipótesis a partir de unos datos experimentales.
5. Trabajar con las distribuciones adecuadas según las muestras utilizadas y los parámetros a estimar.
6. Contrastar hipótesis probabilísticas.

Actividad 1.26. Una *muestra sin reemplazo* es cualquier subconjunto de una población (un ejemplo típico es una mano de cartas de una baraja). Una *muestra con reemplazo* se obtiene escogiendo sucesivamente un determinado número de elementos de la población sin quitarlos de la misma, de forma que pueden repetirse (un ejemplo típico es el resultado de tiradas sucesivas de un dado). Llamaremos *estadístico* a cualquier parámetro poblacional restringido a una muestra. Para distinguirlo del correspondiente parámetro sobre la población, utilizaremos una nomenclatura diferente; así, designaremos la media de una variable aleatoria X en una muestra por $\overline{X}$, y su desviación típica por s(X).
Trabajaremos con distribuciones en 3 ámbitos: en la población, en una muestra y en el conjunto de todas las muestras. Naturalmente, para poder hacer estimaciones sobre una población a partir de una muestra necesitaremos saber cómo se distribuyen los valores del estadístico correspondiente en el conjunto de todas las muestras de la población de un determinado tipo (con o sin reemplazo) y de un determinado tamaño; a esta distribución la llamaremos *distribución muestral*. Las principales propiedades de ésta se resumen en la siguiente tabla, dónde indicamos por n(U) el tamaño de la población y por n el tamaño de la muestra:

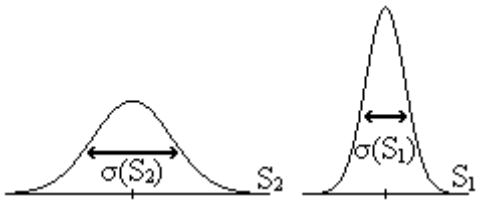
parámetro poblacional Ω	estadístico S	distribución muestral sin reemplazo $\mu(S)$, $\sigma(S)$	distribución muestral con reemplazo $\mu(S)$, $\sigma(S)$
$\mu(X)$	$\overline{X}$	$\mu(\overline{X}) = \mu(X)$	
		$\sigma(\overline{X})^2 = \sigma(X)^2(n(U)-n)/(n\cdot(n(U)-1))$	$\sigma(\overline{X})^2 = \sigma(X)^2/n$
$\sigma(X)$	s(X)	$\mu(s(X)^2) = \sigma(X)^2\cdot n/(n-1)$ $\sigma(s(X))^2 \approx \sigma(X)^2/(2n)$ si $\geq$ n 100.	

Observamos que si $n(U)=\infty$, entonces la varianza de la distribución muestral de medias con y sin reemplazo son iguales. En la práctica, podemos utilizar la fórmula de la distribución muestral con reemplazo si el tamaño n(U) de la población es mucho más

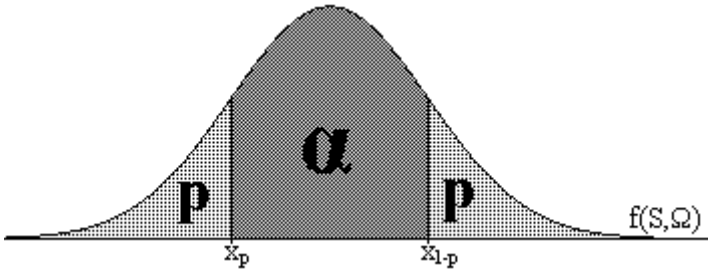
grande que el tamaño n de la muestra. Si no decimos lo contrario, supondremos que éste es el caso.
Problema 1.12: Obtener la varianza de la distribución muestral de medias y la media de la distribución muestral de varianzas con muestras formadas por la repetición 3 veces del lanzamiento de 5 dados anotando en cada lanzamiento el número de ases obtenidos (suponiendo que los dados no están cargados). Dividir la clase en grupos de 3 de modo que cada miembro haga un lanzamiento de 5 dados, calculando en cada grupo la media y la varianza de la muestra obtenida. Calcular la varianza de las medias y la media de las varianzas obtenidas por toda la clase y compararlas con los previos resultados teóricos.

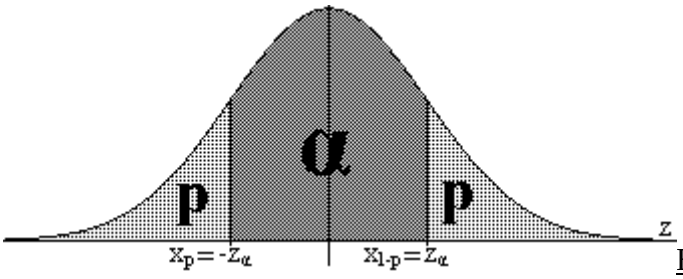
Actividad 1.27. Para estimar correctamente un parámetro poblacional Ω necesitaremos un estadístico S que sea *un estimador insesgado* del mismo, de forma que $\mu(S) = \Omega$. En caso de que no lo sea pero conozcamos el sesgo que se produce, de forma que $\mu(S) = f(\Omega)$, siendo f una función lineal, podemos definir un estimador corregido $\hat{S} = f^{-1}(S)$ tal que $\mu(\hat{S}) = \Omega$.
Ejercicio 1.5: analizar si la media $\overline{X}$ y la varianza s^2 son o no estimadores insesgados de los correspondientes parámetros poblacionales $\mu(X)$ y $\sigma(X)$. En caso de que alguno no lo sea, obtener el correspondiente estadístico corregido y comprobar que es un estimador insesgado.

Actividad 1.28. Si tenemos dos estimadores insesgados S_1 y S_2, diremos que S_1 es más eficiente que S_2 si y solamente si $\sigma(S_1)<\sigma(S_2)$.
Ejercicio 1.6: queremos estimar la media μ de una población a partir de las medias $\overline{X}_1$, $\overline{X}_2$ de dos muestras de tamaño respectivo n_1, n_2 tales que $n_1<n_2$. Qué estimador será más eficiente? Demostrarlo.



Actividad 1.29. Diremos que $[\Omega_1,\Omega_2]$ es un *intervalo de confianza* del $100\alpha\%$ para un parámetro poblacional Ω si la probabilidad de que Ω esté dentro de este intervalo es igual a . α. Para determinarlo necesitaremos conocer la distribución muestral de alguna función $f(S,\Omega)$, siendo S el estadístico de una muestra que utilizamos para estimar Ω. En general, buscaremos en esta distribución muestral de densidad probabilística dos "picos" de probabilidad p, de forma que el área entre los dos "picos" sea α, tal y como se indica en la figura adjunta. Observamos que, comoquiera que el área bajo la curva es 1, se ha de cumplir $2p+\alpha=1$. Las abcises correspondientes a una determinada área se denominan *coeficientes críticos*. Hay que examinar con cuidado la configuración de la tabla de la distribución y las gráficas que la acompañan para determinar a qué área se refiere cada coeficiente crítico (parte de la izquierda, interior, exterior...) y qué son por tanto los coeficientes tales que $x_p \le f(S,\Omega) \le x_{1-p}$ nos da un intervalo de confianza para Ω del $100\alpha\%$.





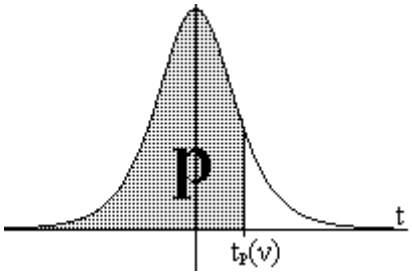
Ejercicio 1.7: si las muestras son grandes ($n\geq 30$) y el parámetro poblacional es la media poblacional, entonces tomando la normalización de la media de la muestra,
$z = f(\overline{X},\mu) = (\overline{X}-\mu)/\sigma(\overline{X})$, se distribuirá aproximadamente de acuerdo con la distribución normal tipificada. Para obtener el intervalo de confianza habremos de calcular primero la media y la desviación típica de la muestra, $\overline{X}$, s; a continuación calcular la desviación típica corregida $\hat{s}$, utilizarla como estimador insesgado de la desviación típica poblacional σ, y a partir del valor estimado de ésta obtener la desviación típica de las medias en la distribución muestral, $\sigma(\overline{X})$. Utilizando la tabla de la distribución normal tipificada (inversa) para obtener el coeficiente crítico z_α tal que la probabilidad de $|z|\leq z_\alpha$ sea α (recordemos que la distribución normal tipificada es simétrica) podremos averiguar el intervalo de confianza para μ. Obtener las fórmulas correspondientes.
Problema 1.13: aplicarlo a la obtención de un intervalo de confianza del 80% para el número mediano de ases resultantes de lanzar 30 veces un dado a partir de los resultados experimentales obtenidos por todos los alumnos de la clase (en un número no inferior a 30).

Actividad 1.30. Si por consideraciones teóricas formulamos la hipótesis de un valor para un parámetro poblacional Ω, y a partir de una muestra experimental obtenemos un intervalo de confianza del 100α% para este parámetro poblacional, si el valor teórico de éste está fuera de este intervalo, es decir
$f(S,\Omega)\notin[x_p,x_{1-p}]$, pueden haber dos explicaciones: la primera es que la teoría y por lo tanto la hipótesis esté equivocada; la segunda es que la muestra sea "anómala", de modo que siendo correcta la teoría el parámetro poblacional Ω esté fuera del intervalo de confianza del 100α%: la probabilidad de esto es $\beta=1-\alpha$. Diremos así que la muestra nos permite rechazar la hipótesis con un *nivel de significación* de β (que será por lo tanto la probabilidad de que nos equivoquemos al rechazar la hipótesis). Naturalmente, solamente podremos rechazar hipótesis con niveles de significación iguales o menores a 0'5, y cuanto menor sea el nivel de significación el rechazo de la hipótesis tendrá más fuerza.
Problema 1.14: ¿con qué nivel de significación podríamos en su caso rechazar la hipótesis de que el dado del Problema 1.13 no está cargado (es decir, que todas las caras del dado tienen la misma probabilidad de salir)?

Actividad 1.31. Si las muestras son pequeñas, su distribución no se aproxima a la normal. Pero si una variable aleatoria X tiene una distribución normal en una población infinita, la distribución del estadístico
$t = f(\overline{X},\mu) = (\overline{X}-\mu(X))/\sigma(\overline{X})$ de las muestras de tamaño n es $Y_\nu(t)=Y_\nu(0)\cdot(1+t^2/\nu)^{-(\nu+1)/2}$ con $\nu=n-1$, que se denomina *distribución **t** de "Student"* con ν grados de libertad. $Y\nu(0)$ se escoge de modo que $\int_{-\infty}^{+\infty} Y_\nu(t)dt=1$.

Teniendo en cuenta que $e = \lim_{u\to\infty}(1+1/u)$, demostrar el
Teorema 1.30: $\lim_{v\to\infty} Y_v(t) = P^{N(0,1)}(t)$ (es decir, la distribución **t** de "*Student*" se aproxima a la distribución normal tipificada cuando el número de grados de libertad se hace muy grande); ¿cuanto valdrá $Y_\infty(0)$?

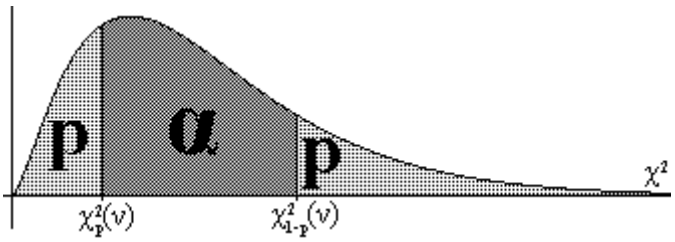
Actividad 1.32. Utilizaremos la tabla de la distribución **t** de "*Student*" (inversa) para determinar el coeficiente crítico $t_p(\nu)$ correspondiente al intervalo de confianza del 100α% de la media poblacional μ a partir de la media $\overline{X}$ y la desviación típica s(X) de una muestra de tamaño n, con las fórmulas obtenidas en el Ejercicio 1.7.



Problema 1.15: obtener un intervalo de confianza del 90% para la media de una variable aleatoria en una población infinita con distribución normal a partir de la muestra (302'23, 302'21, 302'23, 302'22, 302'25).

Actividad 1.33.
Problema 1.16: formando grupos de 3 a 5 estudiantes, cada estudiante en cada grupo deberá lanzar 30 veces un dado y anotar el número de ases obtenidos; hacer estimaciones alrededor de cada dado a partir de la muestra dada por los resultados obtenidos por cada grupo.

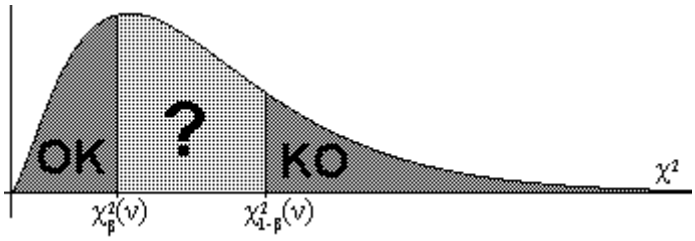
Actividad 1.34. Si una variable aleatoria X tiene una distribución normal en una población infinita, la distribución del estadístico $\chi^2 = f(s,\sigma) = n\cdot s(X)^2/\sigma(X)^2$ de las muestras de tamaño n entre 0 y ∞ es $V_\nu(\chi^2)=K_\nu\cdot(\chi^2)^{(\nu-2)/2}\cdot e^{-\chi 2/2}$ con ν=n-1, que se denomina *distribución Ji-cuadrado* con ν grados de libertad. K_ν se escoge de modo que $\int_0^\infty V_\nu(\chi^2)=1$. Utilizaremos la tabla de la distribución Ji-cuadrado (inversa) para determinar los coeficientes críticos $\chi^2_p(\nu)$ correspondientes al intervalo de confianza del 100α% de la desviación típica poblacional σ a partir de la desviación típica s(X) de una muestra de tamaño n, de modo que $\chi^2_p(\nu) \le \chi^2 \le \chi^2_{1-p}(\nu)$. Obtener la expresión para el intervalo de confianza de la desviación típica poblacional σ(X). Observemos que la desviación típica corregida $\hat{s}(X)$ de la muestra ha de estar necesariamente dentro de este intervalo, comoquiera que es un estimador insesgado de la desviación típica poblacional.



Problema 1.17: obtener un intervalo de confianza del 90% para la desviación típica de una variable aleatoria en una población infinita con distribución normal a partir de la muestra (302'23, 302'21, 302'23, 302'22, 302'25); comprobar que la desviación típica corregida de la muestra está dentro de este intervalo.

Actividad 1.35: Si tenemos un conjunto de k sucesos mutuament excluyentes E_i a los que suponemos una probabilidad $p(E_i)$ para i=1...k, en n ocasiones la frecuencia esperada de cada uno de ellos será respectivamente $e_i=n\cdot p(E_i)$, correspondiente a la

media obtenida en el Teorema 1.16. Si en una muestra de estas n ocasiones las frecuencias observadas son respectivamente o_i, siendo $n \geq 30$ y cumpliéndose $e_i \geq 5$ para todos los sucesos, entonces el estadístico $\chi^2 = \sum_{i=1}^{k} (o_i - e_i)^2/e_i$ se distribuirá aproximadamente de acuerdo con la distribución Ji-cuadrado con $\nu = k-1$ grados de libertad. Si para algún suceso fuera $e_i < 5$ habríamos de agregar sucesos hasta conseguir que se cumpla la condición.



Podemos utilizar este estadístico para estimar la concordancia entre la hipótesis probabilística y los resultados experimentales obtenidos en la muestra. Naturalmente, cuanto menor sea χ^2 habrá una mayor concordancia: diremos que hay *buena concordancia* entre la muestra y la hipótesis probabilística (y por lo tanto aceptamos ésta) con un nivel de significación de β si $\chi^2 < \chi^2_\beta(\nu)$; por el contrario, si $\chi^2_{1-\beta}(\nu) < \chi^2$ podremos *rechazar* la hipótesis probabilística con un nivel de significación de β (que será de nuevo la probabilidad de equivocarnos al rechazarla, es decir la probabilidad de que la hipótesis sea correcta pero hayamos encontrado una muestra entre el 100β% de las muestras más desviadas de las frecuencias medias esperadas); finalmente si $\chi^2_\beta(\nu) \leq \chi^2 \leq \chi^2_{1-\beta}(\nu)$ diremos que los resultados experimentales no son decisivos con este nivel de significación para aceptar o rechazar la hipótesis probabilística. Observamos que una hipótesis probabilística puede ser aceptada (o rechazada) con un nivel de significación "débil" y los resultados no ser decisivos con un nivel de significación más fuerte. Lo que no puede pasar es que con un nivel de significación aceptemos una hipótesis y con otro nivel de significación la rechacemos. Naturalmente, el nivel de significación más débil que podemos utilizar es el de $\beta = 0'5$: si $\chi^2 < \chi^2_{0'5}(\nu)$ tendremos tendencia a aceptar la hipótesis con un nivel de significación mayor o menor, y si $\chi^2 > \chi^2_{0'5}(\nu)$ tendremos tendencia a rechazarla.

Problema 1.18: contrastar la hipótesis de que un dado no está cargado (que todas las caras tienen la misma probabilidad de salir) lanzándolo 30 veces y anotando el número de veces que sale cada cara.

Trabajo 2 (para su realización en equipo)**:**
En 100000 tiradas de 5 dados se obtiene 10 repóqueres, 300 póqueres, 3342 tríos, 16030 parejas y 40198 simples ases. ¿Se podría acusar que los dados están trucados? ¿Con qué nivel de significación en tal caso?