1.3. Fer estimacions sobre una població a partir d'una mostra:

Objectius:

1. Estudiar les propietats de les distribucions de les mostres d'una població.
2. Identificar els estadístics-paràmetres d'una mostra que millor permeten estimar els paràmetres de la població.
3. Construir intervals que continguen amb una certa probabilitat el valor d'un paràmetre poblacional.
4. Determinar la probabilitat d'equivocar-nos al rebutjar una hipòtesi a partir d'unes dades experimentals.
5. Treballar amb les distribucions adequades segons les mostres utilitzades i els paràmetres a estimar.
6. Contrastar hipòtesis probabilístiques.

Activitat 1.26. Una *mostra sense reemplaçament* és qualsevol subconjunt d'una població (un exemple típic és una mà de cartes d'una baralla). Una *mostra amb reemplaçament* s'obté escollint successivament un determinat número d'elements de la població sense llevar-los de la mateixa, de manera que poden repetir-se (en exemple típic és el resultat de tirades successives d'un dau). Anomenarem *estadístic* a qualsevol paràmetre poblacional restringit a una mostra. per a distingir-lo del corresponent paràmetre sobre la població, utilitzarem una nomenclatura diferent; així, designarem la mitjana d'una variable aleatòria X en una mostra per $\overline{X}$, i la seua desviació típica per s(X).
Treballarem amb distribucions en 3 àmbits: en la població, en una mostra i en el conjunt de totes les mostres. Naturalment, per poder fer estimacions sobre una població a partir d'una mostra necessitarem saber com es distribueixen els valors de l'estadístic corresponent en el conjunt de totes les mostres de la població d'un determinat tipus (amb o sense reemplaçament) i d'una determinada grandària; a aquesta distribució l'anomenarem *distribució mostral*. Les principals propietats d'aquesta es resumeixen en la següent tabla, on indiquem per n(U) la grandària de la població i per n la grandària de la mostra:

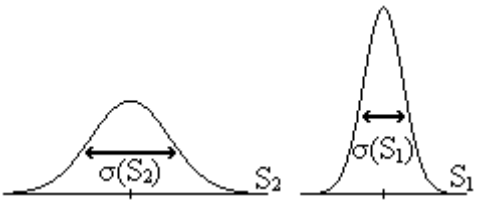
<table>
<tr><td>paràmetre poblacional
Ω</td><td>estadístic
S</td><td>distribució mostral sense reemplaçament
$\mu(S)$, $\sigma(S)$</td><td>distribució mostral amb reemplaçament
$\mu(S)$, $\sigma(S)$</td></tr>
<tr><td rowspan="2">$\mu(X)$</td><td rowspan="2">$\overline{X}$</td><td colspan="2">$\mu(\overline{X}) = \mu(X)$</td></tr>
<tr><td>$\sigma(\overline{X})^2 = \sigma(X)^2(n(U)-n)/(n\cdot(n(U)-1))$</td><td>$\sigma(\overline{X})^2 = \sigma(X)^2/n$</td></tr>
<tr><td>$\sigma(X)$</td><td>s(X)</td><td colspan="2">$\mu(s(X)^2) = \sigma(X)^2\cdot n/(n-1)$
$\sigma(s(X))^2 \approx \sigma(X)^2/(2n)$ si $n\geq 100$</td></tr>
</table>

Observem que si $n(U)=\infty$, aleshores la variança de la distribució mostral de mitjanes amb i sense reemplaçament són iguals. En la pràctica, podem utilitzar la fòrmula de la distribució mostral amb reemplaçament si la grandària n(U) de la població és molt més

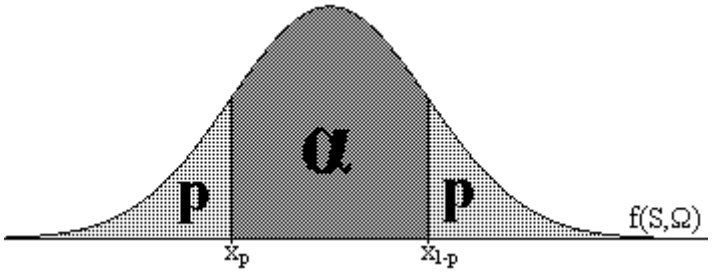
gran que la grandària n de la mostra. Si no diem el contrari, suposarem que aquest és el cas.
Problema 1.12: Obtenir la variança de la distribució mostral de mitjanes i la mitjana de la distribució mostral de variances amb mostres formades per la repetició 3 vegades del llançament de 5 daus anotant en cada llançament el número d'asos obtinguts (suposant que els daus no estan carregats). Dividir la classe en grups de 3 de manera que cada membre faça un llançament de 5 daus, calculant en cada grup la mitjana i la variança de la mostra obtinguda. Calcular la variança de les mitjanes i la mitjana de les variances obtingudes per tota la classe i comparar-les amb els previs resultats teòrics.

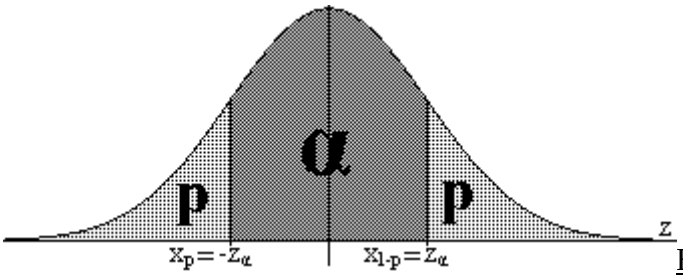
Activitat 1.27. Per a estimar correctament un paràmetre poblacional Ω necessitarem un estadístic S que siga un *estimador inesbiaixat* del mateix, de manera que $\mu(S) = \Omega$. En cas que no ho siga però conegam l'esbiaixament que es produeix, de manera que $\mu(S) = f(\Omega)$, essent f una funció lineal, podem definir un estimador corregit $\hat{S} = f^{-1}(S)$ tal que $\mu(\hat{S}) = \Omega$.
Exercici 1.5: analitzar si la mitjana $\overline{X}$ i la variança s^2 són o no estimadors inesbiaixats dels corresponents paràmetres poblacionals $\mu(X)$ i $\sigma(X)$. En cas que algú no ho siga, obtenir el corresponent estadístic corregit i comprovar que és un estimador inesbiaixat.

Activitat 1.28. Si tenim dos estimadors inesbiaixats S_1 i S_2, direm que S_1 és més eficient que S_2 si i solament si $\sigma(S_1)<\sigma(S_2)$.
Exercici 1.6: volem estimar la mitjana μ d'una població a partir de les mitjanes $\overline{X}_1$, $\overline{X}_2$ de dues mostres de grandària respectiva n_1, n_2 tals que $n_1<n_2$. Quin estimador serà més eficient? Demostrar-ho.



Activitat 1.29. Direm que $[\Omega_1,\Omega_2]$ es un *interval de confiança* del $100\alpha\%$ per a un paràmetre poblacional Ω si la probabilitat de que Ω estiga dins d'aquest interval és igual a α. Per determinar-ho necessitarem conèixer la distribució mostral d'alguna funció $f(S,\Omega)$, essent S l'estadístic d'una mostra que utilitzem per estimar Ω. En general, buscarem en aquesta distribució mostral de densitat probabilística dos "pics" de probabilitat p, de manera que l'àrea entre els dos "pics" siga α, tal com s'indica en la figura adjunta. Observem que, per tal com l'àrea baix de la corva és 1, s'ha d'acomplir $2p+\alpha=1$. Les abcises corresponents a una determinada àrea s'anomenen *coeficients crítics*. Cal examinar amb cura la configuració de la tabla de la distribució i les gràfiques que l'acompanyen per determinar a quina àrea es refereix cada coeficient crític (part de l'esquerra, interior, exterior...) i quins són per tant els coeficients tals que $x_p \leq f(S,\Omega) \leq x_{1-p}$ ens dóna un interval de confiança per a Ω del $100\alpha\%$.





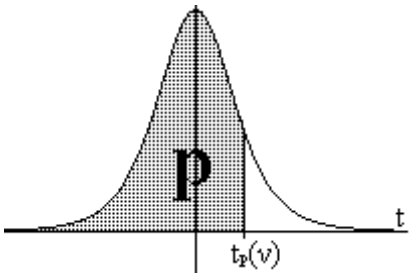
Exercici 1.7: si les mostres són grans ($n\geq 30$) i el paràmetre poblacional és la mitjana poblacional, aleshores prenent la normalització de la mitjana de la mostra,
$z = f(\overline{X},\mu) = (\overline{X}-\mu)/\sigma(\overline{X})$, es distribuirà aproximadament d'acord amb la distribució normal tipificada. Per obtenir l'interval de confiança haurem de calcular primer la mitjana i la desviació típica de la mostra, $\overline{X}$, s; a continuació calcular la desviació típica corregida $\hat{s}$, utilitzar-la com estimador inesbiaxat de la desviació típica poblacional σ, i a partir del valor estimat d'aquesta obtenir la desviació típica de les mitjanes en la distribució mostral, $\sigma(\overline{X})$. Utilitzant la tabla de la distribució normal tipificada (inversa) per obtenir el coeficient crític z_α tal que la probabilitat de $|z|\leq z_\alpha$ siga α (recordem que la distribució normal tipificada és simètrica) podrem averiguar l'interval de confiança per a μ. Obtenir les fòrmules corresponents.
Problema 1.13: aplicar-ho a l'obtenció d'un interval de confiança del 80% per al número mitjà d'asos resultants de llançar 30 vegades un dau a partir dels resultats experimentals obtinguts per tots els alumnes de la classe (en un número no inferior a 30).

Activitat 1.30. Si per consideracions teòriques formulem la hipòtesi d'un valor per a un paràmetre poblacional Ω, i a partir d'una mostra experimental obtenim un interval de confiança del 100α% per a aquest paràmetre poblacional, si el valor teòric d'aquest està fora d'aquest interval, és a dir
$f(S,\Omega)\notin[x_p,x_{1-p}]$, poden haver dues explicacions: la primera és que la teoria i per tant la hipòtesi estiga equivocada; la segona és que la mostra siga "anòmala", de manera que essent correcta la teoria el paràmetro poblacional Ω estiga fora de l'interval de confiança del 100α%: la probabilitat d'això és β=1-α. Direm així que la mostra ens permet rebutjar la hipòtesi amb un *nivell de significació* de β (que serà per tant la probabilitat de que s'equivoquem al rebutjar la hipòtesi). Naturalment, solament podrem rebutjar hipòtesis amb nivells de significació iguals o menors a 0'5, i quant menor siga el nivell de significació el rebuig de la hipòtesi tindrà més força.
Problema 1.14: amb quin nivell de significació podriem en el seu cas rebutjar la hipòtesi de que el dau del Problema 1.13 no està carregat (és a dir, que totes les cares del dau tenen la mateixa probabilitat de sortir)?

Activitat 1.31. Si les mostres són xicotetes, la seua distribució no s'aproxima a la normal. Però si una variable aleatòria X té una distribució normal en una població infinita, la distribució de l'estadístic
$t = f(\overline{X},\mu) = (\overline{X}-\mu(X))/\sigma(\overline{X})$ de les mostres de grandària n és $Y_\nu(t)=Y_\nu(0)\cdot(1+t^2/\nu)^{-(\nu+1)/2}$ amb ν=n-1, que s'anomena *distribució* ***t*** *de "Student"* amb ν graus de llibertat. Yν(0) s'escollís de manera que $\int_{-\infty}^{+\infty} Y_\nu(t)dt=1$.
Tenint en compte que $e = \lim_{u\to\infty}(1+1/u)$, demostrar el
Teorema 1.30: $\lim_{\nu\to\infty} Y_\nu(t) = P^{N(0,1)}(t)$ (és a dir, la distribució **t** de "*Student*" s'aproxima

a la distribució normal tipificada quan el número de graus de llibertat es fa molt gran); quan valdrà $Y_\infty(0)$?

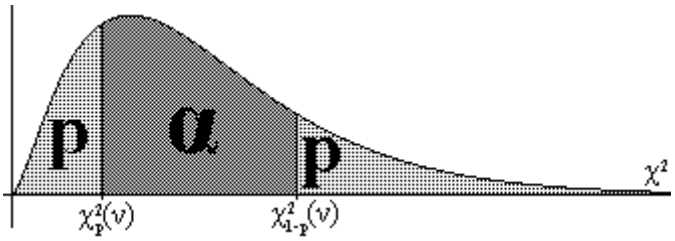
Activitat 1.32. Utilitzarem la tabla de la distribució **t** de "*Student*" (inversa) per a determinar el coeficient crític $t_p(\nu)$ corresponent a l'interval de confiança del 100α% de la mitjana poblacional μ a partir de la mitjana $\overline{X}$i la desviació típica s(X) d'una mostra de grandària n, amb les fòrmules obtingudes en l'Exercici 1.7.



Problema 1.15: obtenir un interval de confiança del 90% per a la mitjana d'una variable aleatòria en una població infinita amb distribució normal a partir de la mostra (302'23, 302'21, 302'23, 302'22, 302'25).

Activitat 1.33.

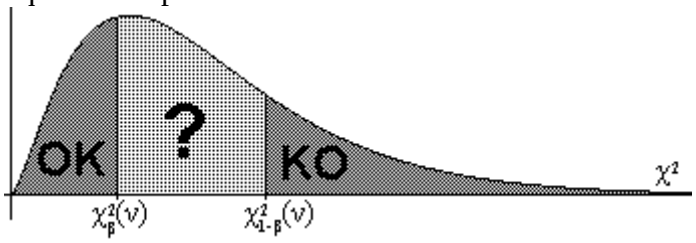
Problema 1.16: formant grups de 3 a 5 estudiants, cada estudiant en cada grup haurà de llançar 30 vegades un dau i anotar el número d'asos obtinguts; fer estimacions al voltant de cada dau a partir de la mostra donada pels resultats obtinguts per cada grup.



Activitat 1.34. Si una variable aleatòria X té una distribució normal en una població infinita, la distribució de l'estadístic $\chi^2 = f(s,\sigma) = n\cdot s(X)^2/\sigma(X)^2$ de les mostres de grandària n entre 0 i ∞ és $V_\nu(\chi^2)=K_\nu\cdot(\chi^2)^{(\nu-2)/2}\cdot e^{-\chi 2/2}$ amb ν=n-1, que s'anomena *distribució Xi-quadrat* amb ν graus de llibertat. K_ν s'escollís de manera que $\int_0^\infty V_\nu(\chi^2)=1$. Utilitzarem la tabla de la distribució Xi-quadrat (inversa) per a determinar els coeficient crítics $\chi^2_p(\nu)$ corresponents a l'interval de confiança del 100α% de la desviació típica poblacional σ a partir de la desviació típica s(X) d'una mostra de grandària n, de manera que $\chi^2_p(\nu) \le \chi^2 \le \chi^2_{1-p}(\nu)$. Obtenir l'expressió per a l'interval de confiança de la desviació típica poblacional σ(X). Observem que la desviació típica corregida $\hat{s}(X)$ de la mostra ha d'estar necessàriament dins d'aquest interval, per tal com és un estimador inesbiaixat de la desviació típica poblacional.

Problema 1.17: obtenir un interval de confiança del 90% per a la desviació típica d'una variable aleatòria en una població infinita amb distribució normal a partir de la mostra (302'23, 302'21, 302'23, 302'22, 302'25); comprovar que la desviació típica corregida de la mostra està dins d'aquest interval.

Activitat 1.35: Si tenim un conjunt de k successos mutuament excloents E_i als que suposem una probabilitat $p(E_i)$ per a i=1...k, en n ocasions la freqüència esperada de cadascú d'ells serà respectivament $e_i=n\cdot p(E_i)$, corresponent a la mitjana obtinguda en el Teorema 1.16. Si en una mostra d'aquestes n ocasions les freqüències observades són respectivament oi, essent $n\geq 30$ i acomplint-se $e_i\geq 5$ per a tots els successos, aleshores l'estadístic $\chi^2 = \sum_{i=1}^{k} (o_i-e_i)^2/e_i$ es distribuirà aproximadament d'acord amb la distribució Xi-quadrat amb $\nu=k-1$ graus de llibertat. Si per a algun succés fora $e_i<5$ hauriem d'agregar successos fins aconseguir que s'acomplisca la condició.



Podem utilitzar aquest estadístic per estimar la concordància entre la hipòtesi probabilística i els resultats experimentals obtinguts en la mostra. Naturalment, quan menor siga χ^2 hi haurà una major concordància: direm que hi ha *bona concordància* entre la mostra i la hipòtesi probabilística (i per tant acceptem aquesta) amb un nivell de significació de β si $\chi^2<\chi^2_{\beta}(\nu)$; pel contrari, si $\chi^2_{1-\beta}(\nu)<\chi^2$ podrem *rebutjar* la hipòtesi probabilística amb un nivell de significació de β (que serà de nou la probabilitat d'equivocar-nos al rebutjar-la, és a dir la probabilitat de que la hipòtesi siga correcta però hagem trobat una mostra entre el 100β% de les mostres més desviades de les freqüències mitjanes esperades); finalment si $\chi^2_{\beta}(\nu)\leq\chi^2\leq\chi^2_{1-\beta}(\nu)$ direm que els resultats experimentals no son decisius amb aquest nivell de significació per a acceptar o rebutjar la hipòtesi probabilística. Observem que una hipòtesi probabilística pot ser acceptada (o rebutjada) amb un nivell de significació "feble" i els resultats no ser decisius amb un nivell de significació més fort. El que no pot passar és que amb un nivell de significació acceptem una hipòtesi i amb altre nivell de significació la rebutgem. Naturalment, el nivell de significació més feble que podem utilitzar és el de β=0'5: si $\chi^2<\chi^2_{0'5}(\nu)$ tindrem tendència a acceptar la hipòtesis amb un nivell de significació major o menor, i si $\chi^2>\chi^2_{0'5}(\nu)$ tindrem tendència a rebutjar-la.

Problema 1.18: contrastar la hipòtesi de que un dau no està carregat (que totes les cares tenen la mateixa probabilitat de sortir) llançant-lo 30 vegades i anotant el número de vegades que surt cada cara.

Treball 2 (per a la seua realització en equip)**:**
En 100000 tirades de 5 daus s'obté 10 repoquer, 300 pòquers, 3342 trios, 16030 parelles i 40198 simples asos. Es podria acusar que els daus estan trucats? Amb quin nivell de significació, en tal cas?