Aproximarea distributiei binomiale cu o distributie normala. Aceasta înseamna ca putem exprima valorile ***z*** în termeni de N, P si Q. Formula originala pentru ***z*** ne amintim ca este:

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

din care, prin substituire, se construieste formula pentru ***z*** binomial:

$$z = \frac{X - N * P}{\sqrt{N * P * Q}}$$

$$z^2 = \frac{(X - N * P)^2}{N * P * Q}$$

Daca, înainte de ridicarea la patrat, ***z*** urmeaza o distributie normala, dupa ridicarea la patrat z urmeaza un alt tip de distributie, „*chi-patrat*”

Formula de calcul este una derivata din testul ***z***:

$$\chi^2 = \sum \frac{(f_O - f_E)^2}{f_E}$$

unde f_O este frecventa observata iar f_E, frecventa asteptata.

Chi-patrat – pentru gradul de corespondenta (*Goodness of Fit*)

Aceasta varianta a testului chi-patrat compara frecventele observate ale unei distributii cu frecventele teoretice (asteptate) ale acelei variabile. De exemplu, daca avem frecventele unei variabile putem afla daca aceasta se distribuie dupa curba normala (z), prin compararea cu frecventele cunoscute ale acestei distributii (aria de sub curba).

Sa presupunem ca a fost aplicat un test de cunostinte unui esantion de 200 de elevi, care a fost evaluat cu calificative, astfel: *F.Slab, Slab, Mediu, Bun, F.Bun*.

Problema cercetarii: Calificativele obtinute se distribuie normal la nivelul clasei?

Populatia 1: Calificativele obtinute de elevi.

Populatia 2: Calificativele asa cum s-ar distribui pe o curba normala: FS=2.5%, B=14%, M=67%, B=14% si FB=2.5% (procentele sunt cele tipice unei curbe **z**, împartite în cinci clase valorice)

Ipoteza de nul (H_0): Distributia calificativelor este aceeasi ca în cazul curbei normale.

Ipoteza cercetarii (H1): Distributia calificativelor clasei este diferita de distributia normala (exprimând speranta cercetatorului de a avea mai multe calificative spre zona superioara a distributiei).

Determinarea caracteristicilor deciziei statistice:

- alegem α=0.05 (în cazul testului χ^2 decizia nu poate fi decât unilaterala, deoarece acest test nu poate lua valori negative)

- gasim valoarea critica pentru χ^2 =9.48 în tabela pentru distributia χ^2, pentru df=(2-1)*(5-1)=4 si α=0.05

Tabelul urmator contine datele de cercetare si algoritmul de calcul:

Calificativ	**Frecventa observata** **(f_O)**	**Frecventa asteptata** **(f_E)**	$\frac{(f_O - f_E)^2}{f_E}$
FB	10	2.5% of 200 =5	$\frac{(10-5)^2}{5} = 5.00$
B	34	14% of 200 =28	$\frac{(34-28)^2}{28} = 1.29$
M	140	67% of 200 =134	$\frac{(140-134)^2}{134} = 0.27$
S	10	14% of 200 =28	$\frac{(10-28)^2}{28} = 11.57$
FS	6	2.5% of 200 =5	$\frac{(6-5)^2}{5} = 0.20$
Σ	200	-	$\chi^2 = \sum \frac{(f_O - f_E)^2}{f_E} = 18\ 33$

Decizia statistica:

- χ^2 calculat (18,33) este mai mare decât χ^2 critic (9,48)

- Respingem ipoteza de nul si tragem concluzia ca distributia calificativelor nu urmeaza forma curbei normale.

Testul z pentru diferenta dintre proportiile a doua esantioane independente

$$z = \frac{p - P}{\sqrt{\frac{PQ}{N}}}$$

Un studiu pe doua esantioane din doua tari diferite conduce la constatarea ca proportia (p_1=0.15) stângacilor a esantionului (n_1=100) dintr-o tara este diferita de proportia (p_2=0.25) stângacilor din esantionul corespunzator celeilalte tari (n_2=90). Este firesc sa ne punem întrebarea daca exista într-adevar o diferenta dintre proportia stângacilor din cele doua tari (pe care o vom nota cu litere mari: P_1 respectiv P_2) sau daca, dimpotriva, diferentele constatate sunt doar expresia variabilitatii de esantionare.

În acest caz:

- ipoteza cercetarii sustine ca proportiile la nivelul populatiilor sunt diferite ($P_1 \neq P_2$)

- ipoteza de nul sustine ca proportiile celor doua populatii sunt identice ($P_1=P_2$) si, deci, ca diferenta lor este 0 ($P_1-P_2=0$)

În exemplul nostru, P_1 si P_2 reprezinta probabilitatile unui eveniment aleator de tip binomial, în care evenimentul complementar (Q_1, respectiv Q_2) este caracteristica de a fi „dreptaci” (vom ignora acum faptul ca pot exista si „ambidextri”).

Distributia ipotezei de nul pentru diferentele dintre cele doua proportii este aproximata de distributia normala z. Testul statistic va urma modelul testului pentru diferenta dintre mediile a doua esantioane independente:

$$z = \frac{(p_1 - p_2) - (P_1 - P_2)}{\sigma_{(p_1 - p_2)}}$$

unde:

p_1 si p_2 sunt proportiile evenimentului la nivelul esantioanelor

P_1 si P_2 sunt proportiile evenimentului la nivelul populatiei

$\sigma_{(p1-p2)}$ este eroarea standard a distributiei de esantionare

Având în vedere ipoteza de nul ($P_1-P_2=0$), rezulta ca la numitor se va pastra doar diferenta dintre proportiile esantioanelor (p_1-p_2).

La rândul ei, eroarea standard de esantionare a diferentei proportiilor se calculeaza astfel:

$$\sigma_{(p_1-p_2)} = \sqrt{\frac{p_1 * q_1}{n_1} + \frac{p_2 * q_2}{n_2}}$$

unde:

q1 si q2 sunt proportiile complementare ale lui p_1, respectiv p_2 ($q_1=1-p_1$, respectiv $q_2=1-p_2$)

n_1 si n_2 sunt volumele celor doua esantioane

Ca urmare, formula pentru testul diferentei dintre proportiile a doua esantioane independente devine:

$$z = \frac{p_1 - p_2}{\sqrt{\frac{p_1 * q_1}{n_1} + \frac{p_2 * q_2}{n_2}}}$$

Aceasta formula este adecvata atunci când esantioanele sunt suficient de mari (>30). În caz contrar, numaratorul formulei suporta o corectie (***corectia Yates***), dupa cum urmeaza:

$$z = \frac{\left(p_1 - \frac{1}{2 * n_1}\right) - \left(p_2 - \frac{1}{2 * n_2}\right)}{\sqrt{\frac{p_1 * q_1}{n_1} + \frac{p_2 * q_2}{n_2}}}$$

Daca privim graficele distributiilor binomiale ,vom observa ca, spre deosebire de curba normala z, acestea au un caracter „discontinuu”, cu treceri în „trepte” la o valoare la alta. Din acest motiv se recomanda aplicarea unei ***„corectii de continuitate***”, prin scaderea valorii 0.5 din valoarea numaratorului, luata în sens absolut.

Pentru exemplul nostru, vom utiliza formula de mai sus fară corecție :

$$z = \frac{0.15 - 0.25}{\sqrt{\frac{0.15 * 0.85}{100} + \frac{0.25 * 0.75}{90}}} = \frac{-0.10}{\sqrt{0.001 + 0.002}} = \frac{-0.10}{0.054} = -1\,85$$

Daca ne-am propus un test bilateral la un nivel alfa=0.05 (pentru care z critic pe curba normala este egal cu 1.96), atunci va trebui sa acceptam ipoteza de nul si sa concluzionam ca nu se confirma existenta unei diferente semnificative între proportia stângacilor din cele doua comunitati.