Până la sfârșitul secolului XIX, distribuția normală a fost considerată legea universală a variației datelor. Cu toate acestea, K. Pearson a remarcat că frecvențele empirice pot fi foarte diferite de distribuția normală. Întrebarea era cum să dovedești asta. Nu a fost necesară doar o comparație grafică, care este de natură subiectivă, ci și o justificare cantitativă strictă.

Deci a fost inventat criteriul χ2 (pătratul chi), care verifică semnificația discrepanței dintre frecvențele empirice (observate) și teoretice (așteptate). Acest lucru s-a întâmplat încă din 1900, dar criteriul este încă în mișcare astăzi. Mai mult, a fost adaptat pentru a rezolva o gamă largă de probleme. În primul rând, este o analiză a datelor categorice, adică. cele care sunt exprimate nu prin cantitate, ci prin apartenența la o anumită categorie. De exemplu, clasa mașinii, sexul participantului la experiment, tipul de plantă etc. Operațiunile matematice, cum ar fi adunarea și înmulțirea, nu pot fi aplicate la astfel de date, numai pentru ei poate fi calculată frecvența.

Frecvențele observate sunt notate cu **O (observat),** așteptat - **E (așteptat).** Ca exemplu, luați rezultatul unei aruncări de **60 de ori** a unui zar. Dacă este simetrică și omogenă, probabilitatea de a cădea din ambele părți este **1/6** și, prin urmare, cantitatea preconizată de cădere din fiecare parte este de **10 (1/6 1/ 60).** Scriem frecvențele observate și așteptate într-un tabel și desenăm o histogramă.

Frecvențele observate sunt notate cu **O (observat),** așteptat - E **(așteptat).** Ca exemplu, luați rezultatul unei aruncări **de 60** de ori a unui zar. Dacă este simetrică și omogenă, probabilitatea de a cădea din ambele părți este **1/6** și, prin urmare, cantitatea preconizată de cădere din fiecare parte este de **10 (1/6 1/ 60).** Scriem frecvențele observate și așteptate într-un tabel și desenăm o histogramă.

Ipoteza nulă este că frecvențele sunt consecvente, adică datele reale nu contrazic așteptările. O ipoteză alternativă este aceea că abaterile frecvențelor depășesc oscilațiile aleatorii, discrepanțele sunt semnificative statistic. Pentru a face o concluzie riguroasă, avem nevoie:

1. O măsură generalizată a discrepanței dintre frecvențele observate și cele așteptate.

2. Distribuirea acestei măsuri cu validitatea ipotezei potrivit căreia nu există diferențe.

Să începem cu distanța dintre frecvențe. Dacă luăm doar diferența O-E, atunci o astfel de măsură va depinde de scara datelor (frecvențe). De exemplu, **20 - 5 = 15** și **1020 - 1005 = 15**. În ambele cazuri, diferența este 15. Dar, în primul caz, frecvențele preconizate sunt de 3 ori mai mici decât cele observate, iar în al doilea caz, doar 1,5%. Este necesară o măsură relativă, independentă de scară.

Să fim atenți la următoarele fapte. În cazul general, numărul de categorii cu care se măsoară frecvențele poate fi mult mai mare, astfel încât probabilitatea ca o singură observație să se încadreze într-o altă categorie este destul de mică. Dacă da, atunci distribuirea unei astfel de variabile aleatorii va respecta legea evenimentelor rare, cunoscută sub numele de legea Poisson. În legea Poisson, după cum se știe, valorile așteptării și variației matematice coincid (parametrul λ). Aceasta înseamnă că frecvența preconizată pentru o anumită categorie a variabilei nominale Ei va fi în același timp dispersia acesteia. În plus, cu un număr mare de observații, legea lui Poisson tinde să fie normală.

Combinând aceste două fapte, descoperim că dacă ipoteza de acord între frecvențele observate și cele așteptate este adevărată, atunci, cu un număr mare de observații, expresia

$$\frac{O_i - E_i}{\sqrt{E_i}}$$

are o distribuție normală standard.

Este important să ne amintim că normalitatea se va manifesta numai la frecvențe suficient de înalte. În statistici, se acceptă, în general, că numărul total de observații (suma frecvențelor) trebuie să fie de cel puțin 50 și frecvența preconizată în fiecare gradare să fie de cel puțin 5. Doar în acest caz, valoarea prezentată mai sus are o distribuție normală standard. Să presupunem că această condiție este îndeplinită.

În distribuția normală standard, aproape toate valorile se încadrează în ± 3 (regula trei sigma). Astfel, am obținut o diferență relativă a frecvențelor pentru o gradație. Avem nevoie de o măsură generală. Este doar imposibil să adăugăm toate abaterile - primim 0 (ghicim de ce). Pearson a sugerat să pliezi pătratele acestor abateri:

$$\chi_n^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Acesta este semnul criteriului Chi-pătrat Pearson. Dacă frecvențele corespund într-adevăr cu cele așteptate, atunci valoarea criteriului va fi relativ mică (deoarece majoritatea abaterilor sunt aproape de zero). Dar dacă criteriul este mare, atunci acesta este în favoarea diferențelor semnificative între frecvențe.

Criteriul „mare" al lui Pearson devine atunci când apariția unei astfel de semnificații sau chiar mai mari devine puțin probabilă. Și pentru a calcula o astfel de probabilitate, este necesar să cunoaștem distribuția criteriului pentru repetarea repetată a experimentului, când ipoteza acordului de frecvență este adevărată.

După cum este ușor de observat, **valoarea chi-pătrat depinde și de numărul de termeni**. Cu cât sunt mai multe, cu atât criteriul ar trebui să fie mai mare, deoarece fiecare termen va contribui la suma totală. **Prin urmare, pentru fiecare număr de termeni independenți, va exista o distribuție proprie. Se dovedește că χ2 este o întreagă familie de distribuții.**

Și aici ajungem într-un moment delicat. Care este numărul de termeni independenți? Se pare că orice termen (adică deviere) este independent. K. Pearson a crezut și el, dar s-a dovedit a fi greșit. **De fapt, numărul termenilor independenți va fi unul mai mic decât numărul de gradații a variabilei nominale n. De ce? Deoarece, dacă avem un eșantion prin care suma frecvențelor a fost deja calculată, atunci una dintre frecvențe poate fi definită întotdeauna ca diferența dintre numărul total și suma tuturor celorlalte. Prin urmare, variația va fi ceva mai mică. Ronald Fisher a observat acest fapt la 20 de ani după ce Pearson și-a dezvoltat criteriul. Până și mesele trebuiau refăcute.**

Cu această ocazie, **Fisher a introdus în statistici un concept nou - gradul de libertate (grade de libertate),** care reprezintă **numărul de termeni independenți în sumă.** Conceptul de grade de

libertate are o explicație matematică și se manifestă doar în distribuțiile legate de normal (Student, Fisher-Snedekor și chi-pătratul în sine).

Pentru a înțelege mai bine sensul gradelor de libertate, apelăm la analogul fizic. Imaginează-ți un punct care se mișcă liber în spațiu. Are 3 grade de libertate, pentru că se poate deplasa în orice direcție a spațiului tridimensional. Dacă punctul se deplasează pe orice suprafață, atunci are deja două grade de libertate (înainte și înapoi, stânga și dreapta), deși continuă să se afle în spațiul tridimensional. Punctul care se mișcă de-a lungul arcului se află din nou în spațiul tridimensional, dar are un singur grad de libertate, deoarece se poate deplasa fie înainte, fie înapoi. După cum puteți vedea, spațiul în care se află obiectul nu corespunde întotdeauna libertății reale de mișcare.

Aproximativ, distribuția criteriului statistic poate depinde de mai puține elemente decât termenii necesari pentru calculul acestuia. În cazul general, numărul de grade de libertate este mai mic decât observațiile în funcție de numărul de dependențe disponibile.

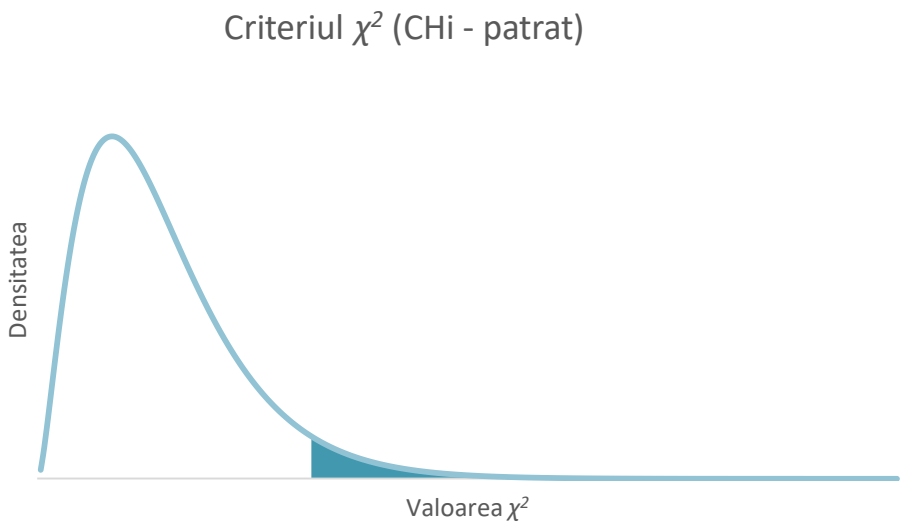
Astfel, distribuția **chi-pătrat (χ2)** este o familie de distribuții, fiecare depinzând de parametrul de grade de libertate. Iar definiția formală a criteriului **chi-pătrat** este următoarea. Distribuția **χ2 (chi-pătrat)** cu k grade de libertate este distribuția sumei pătratelor k variabilelor aleatorii normale standard independente.

În continuare, am putea trece la formula în sine, care calculează funcția de distribuție a **chi-pătratului**, dar, din fericire, totul a fost mult timp calculat pentru noi. Pentru a obține probabilitatea de interes, puteți utiliza fie tabelul statistic corespunzător, fie o funcție gata pregătită în Excel.

Este interesant să vedem cum se modifică forma distribuției **chi-pătrat** în funcție de numărul de grade de libertate.

Test de ipoteză după chi pătrat criteriu Pearson

Așadar, ajungem la testarea ipotezelor folosind metoda chi-pătrat. În general, tehnica rămâne aceeași. Este prezentată o ipoteză nulă conform căreia frecvențele observate corespund celor așteptate (adică nu există nicio diferență între ele, deoarece sunt preluate de la aceeași populație generală). Dacă este așa, atunci răspândirea va fi relativ mică, în cadrul fluctuațiilor aleatorii. Măsura dispersiei este determinată de criteriul chi-pătrat. Mai mult, fie criteriul în sine este comparat cu o valoare critică (pentru nivelul corespunzător de semnificație și grade de libertate), fie, mai corect, se calculează valoarea p observată, adică. probabilitatea obținerii unei astfel de valori sau chiar mai mari a criteriului dacă ipoteza nulă este adevărată.



pentru că ne interesează acordul de frecvențe, atunci respingerea ipotezei va avea loc atunci când criteriul este mai mare decât nivelul critic. Ie criteriul este unilateral. Cu toate acestea, uneori (uneori) este necesară testarea ipotezei din partea stângă. De exemplu, când datele empirice sunt atât de asemănătoare cu cele teoretice. Atunci criteriul poate cădea în regiunea improbabilă, dar deja la stânga. Cert este că, în condiții naturale, este puțin probabil să se obțină frecvențe care coincid practic cu cele teoretice. Există întotdeauna o șansă care dă o eroare. Dar dacă nu există o astfel de eroare, atunci datele au fost falsificate. Dar totuși testează de obicei ipoteza dreaptă.

$$\chi_6^2 = \frac{(8-10)^2}{10} + \frac{(12-10)^2}{10} + \cdots + \frac{(8-10)^2}{10} = 3{,}4$$

Adică, cantilul este 0.05 chi distribuție pătrată (coadă dreaptă) cu 5 grade de libertate χ20.05; 5 = 11.1.

Comparați valorile reale și tabulare. 3,4 (χ2) <11,1 (χ20,05; 5). Criteriul de calcul sa dovedit a fi mai mic, ceea ce înseamnă că ipoteza egalității (acordului) frecvențelor nu este respinsă. În figură, situația arată așa.

Dacă valoarea calculată ar intra în regiunea critică, ipoteza nulă ar fi respinsă.

Va fi mai corect să se calculeze și valoarea **p.** Pentru s-ar face acest lucru, trebuie să găsiți cea mai apropiată valoare pentru un anumit număr de grade de libertate în tabel și să vedeți nivelul de semnificație corespunzător. Dar acesta este secolul trecut. Folosim calculatoare, în special MS Excel. Există mai multe funcții legate de chi-pătrat în Excel.