

Hi-Kvadrat Test

Hi-kvadrat test je statistički test koji se koristi za ispitivanje odnosa između dvije kategoričke varijable. Ovaj test je posebno koristan kada želimo da utvrdimo da li postoji značajna povezanost između varijabli u dvosmjernoj tabeli.

Dvosmjerna Tabela

Dvosmjerna tabela prikazuje odnose između dvije kategoričke varijable. Na primjer, možemo koristiti dvosmjernu tabelu za prikaz odnosa između diplomskog statusa i rase studenata. Da bismo najbolje shvatili informacije sadržane u ovakvoj tabeli, prvo treba pogledati distribuciju svake varijable posebno. Distribucija kategoričke varijable govori koliko se često svaki ishod dogodio.

Izračunavanje Procena

Da bismo opisali odnose između kategoričkih varijabli, izračunavamo odgovarajuće procenete iz datih brojeva. Na primjer, ako želimo da izračunamo procenat bijelaca među studentima, koristimo formulu:

$$\text{Procenat bijelaca} = \left(\frac{10,896}{17,684} \right) \times 100\% = 0.616 \times 100\% = 61.6\%$$

Hi-Kvadrat Statistika

Hi-kvadrat statistika, označena kao χ^2 , je mjera koliko su posmatrani brojevi u dvosmjernoj tabeli udaljeni od očekivanih brojanja. Formula za hi-kvadrat statistiku je:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Primjer Izračuna

Evo broja ispitanika i procenata koji su uspjeli identificirati svih pet činjeničnih tvrdnji kao činjenice. Procenati uzoraka ispitanika koji su svih pet činjeničnih tvrdnji identifikovali kao činjenične su prilično različiti. Konkretno, procenat "veoma pametnih" ispitanika koji su dali svih pet tačnih je bio mnogo veći nego za grupe "donekle pametnih" i "nepametnih". Da li su ovi podaci dobar dokaz da postoji veza između digitalne pameti i sposobnosti da se činjenične izjave identificiraju kao činjenične u populaciji svih odraslih Amerikanaca?

Očekivani broj u bilo kojoj ćeliji dvosmjerne tablice kada je H_0 istinit je:

$$\text{Očekivani broj} = \frac{\text{ukupno redova} \times \text{ukupno kolona}}{\text{ukupno tablica}}$$

Na primjer:

$$\text{Očekivani broj} = \frac{480 \times 260}{1000} = 124.8$$

Izračunavanje hi-kvadrat statistike:

$$\chi^2 = \frac{(168-124.8)^2}{124.8} + \frac{(312-355.2)^2}{355.2} + \frac{(70-91.0)^2}{91.0} + \frac{(280-259.0)^2}{259.0} + \frac{(22-44.2)^2}{44.2} + \frac{(148-125.8)^2}{125.8} = 14.95 + 5.25 + 4.85 + 1.70 + 11.15 + 3.92 = 41.82$$

Hi-Kvadrat Distribucija

Distribucija uzorkovanja hi-kvadrat statistike χ^2 kada je tačna nulta hipoteza da nema povezanosti naziva se hi-kvadrat distribucija. Hi-kvadrat distribucije su familija distribucija koje uzimaju samo nenegativne vrijednosti i nagnute su udesno. Specifična hi-kvadrat distribucija je specificirana davanjem njenih stupnjeva slobode.

Hi-kvadrat test za dvosmjernu tabelu sa r redova i c kolona koristi kritične vrijednosti iz hi-kvadrat distribucije sa $(r-1)(c-1)$ stepenima slobode. Budući da χ^2 mjeri koliko su uočeni brojevi udaljeni od onoga što bi se očekivalo da je H_0 istinit, velike vrijednosti su dokaz protiv H_0 .

Stepeni Slobode

U našem slučaju $r=3$ i $c=2$

$$\text{Stepeni slobode} = (r-1)(c-1) = (3-1)(2-1) = (2)(1) = 2$$

Pogledajte red $df=2$ tabele koju smo naveli. Vidimo da je $\chi^2=41.82$ veći od kritične vrijednosti 9.21 potrebne za značajnost na nivou $\alpha=0.01$. Primjer pokazuje statistički značajan odnos ($P<0.01$) između digitalne "pameti" i sposobnosti da se pet činjeničnih izjava ispravno identifikuje kao činjenice.

Kada Koristiti Hi-Kvadrat Test

Možete bezbedno koristiti hi-kvadrat test kada nije više od 20% očekivanog broja manje od 5 i svi pojedinačni očekivani brojevi su 1 ili veći.

Simpsonov Paradoks

Asocijacija ili poređenje koje vrijedi za sve od nekoliko grupa može nestati ili čak promijeniti smjer kada se podaci kombinuju u jednu grupu. Ova situacija se naziva Simpsonov paradoks.