P_{ij}-valoro: kiel esprimi la malprobablon de la ofto de iu ĉelo en oftotabelo

(P_{ij}-valor: cómo expresar la rareza de la frecuencia de una cierta celda en una tabla de contingencia)

D-ro Carlos Enrique Carleos Artime
Statistika Departemento ĉe Universitato Oviedo
carleos@uniovi.es

Resumo:

Du-dimensia oftotabelo povas esti submetata al statistika testo χ^2 (ĥi-kvadrata) pri nedependeco aŭ homogeneco; tian teston oni povas solvi per la χ^2-adedo. Se oni malakceptas la nulan hipotezon, ofte oni demandas aposteriore tion, kiuj ĉeloj kulpas pri la malakcepto. En la jena artikolo estas traktata la afero esprimi la influon de ĉiu ĉelo al malakcepto de la nula hipotezo kaj estas proponata la P_{ij}-valoro kiel nova esprimmaniero.

Ŝlosilvortoj: oftotabelo, elstara ĉelo, P-valoro.

Resumo:

Una tabla de contingencia de doble entrada puede someterse a un contraste estadístico χ^2 (ji-cuadrado) de independencia u homogeneidad; dicho contraste puede resolverse mediante el estadístico χ^2. Si se rechaza la hipótesis nula, se plantea a menudo a posteriori la pregunta de qué celdas son responsables del rechazo. Se discute en este artículo la cuestión de expresar la influencia de cada celda en el rechazo de la hipótesis nula y se propone el P_{ij}-valor como nueva forma de expresión.

Palabras clave: tabla de contingencia, celda atípica, P-valor.

Enkonduko

En Priskriba aŭ Deskripta Statistiko, oni konsideras muestron, kies anoj havas trajton, statistikan variablon. Nomante X iun variablon, oni priskribas ĝian distribuon per oftotabelo, kiu enhavas almenaŭ:

- La k valorojn kiujn povas preni la variablo: x_1, ..., x_k.
- La respondajn k absolutajn oftojn (n_1, ..., n_k) kaj/aŭ k relativajn oftojn (f_1, ..., f_k), tiel ke

$$f_i = n_i / n \text{ kie } n = \Sigma_{i=1}^{n} n_i$$

Priskribante du-dimensian variablon, ekzemple (X,Y), oni povas indiki la kundistribuon ankaŭ per absolutaj oftoj, tiel ke n_{ij} estas la nombro de muestranoj kie $X=x_i$ kaj $Y=y_j$; aŭ per relativaj oftoj, tiel ke $f_{ij}=n_{ij}/n$.

En Probablokalkulo, anstataŭ statistikajn oni traktas hazardajn variablojn aŭ stokastojn. Diskretaj stokastoj iel similas la statistikajn variablojn, ĉefe se tiuj prenas nur finian nombron de malsamaj valoroj. Se (X,Y) estas du-dimensia diskreta stokasto, ĝia distribuo estas la probabloj $\Pr(X=x_i,Y=y_j)$.

Oni diras ke du diskretaj stokastoj X kaj Y estas nedependaj se $\Pr(X=x_i,Y=y_j) = \Pr(X=x_i)\cdot\Pr(Y=y_j)$. Nedependeco estas grava afero, do oni disvolvis plurajn statistikajn metodojn por konstati ĉu oftotabelo povas respondi al nedependeca distribuo aŭ se, male, ĝi apartiĝas signifikance disde tiu hipotezo.

Mi menciu nun ke kiam oni testas pri homogeneco de distribuo inter pluraj populacioj (tio estas, la variablo havas saman distribuon, sendepende de la populacio), la statistikaj metodoj estas kutime samaj kiel tiuj por testi nedependecon, kvankam la muestrada metodo malsamas (por testi homogenecon, oni muestras en ĉiu populacio difinitan nombron de muestranoj; por testi nedependecon, oni muestras en unu populacio kaj poste oni asignas ĉiun muestranon al grupo laŭ la valoro prenita de iu statistika variablo).

Do, ni pensu pri du diskretaj variabloj, X kaj Y, kaj ni volu testi la hipotez-paron:

H_0: por ĉiu $i = 1, ..., k$; $j = 1, ..., l$; $\Pr(X=x_i, Y=y_j) = \Pr(X=x_i) \cdot \Pr(Y=y_j)$

H_1: ekzistas i, j, tiaj ke $\Pr(X=x_i, Y=y_j) \neq \Pr(X=x_i) \cdot \Pr(Y=y_j)$

Por tia testo ekzistas pluraj adedoj, inter kiuj plej konata estas la χ^2 de Karl Pearson,

$$\chi 2=\Sigma_{i,j}(O_{ij}-E_{ij})^2/E_{ij}$$

kie O_{ij} prezentas la observitan en la muestro valoron por la ĉelo (i,j), tio estas, $O_{ij}=f_{ij}$, kaj E_{ij} prezentas la ekspektikan valoron por tiu sama ĉelo, sub la hipotezo de nedependeco, tio estas, $E_{ij} = n \cdot \Pr(X=x_i) \cdot \Pr(Y=y_j)$. Ĉar relativa ofto estas stimoj de la responda probablo, tiam

$$E_{ij} = n \cdot f_{i\cdot} \cdot f_{\cdot j} = n_{i\cdot} \cdot n_{\cdot j} / n$$

kie

$$n_{i\cdot} = \Sigma_j n_{ij} \text{ kaj } n_{\cdot j} = \Sigma_i n_{ij}$$

kaj simile por la relativaj oftoj.

Se la rezulto de la testo estas teni H_0, tiam finite. Sed se oni malakceptas H_0, kutime ŝprucas la demando: kiuj ĉeloj kulpas pri la malakcepto?

La testo χ^2

Unuan respondon povas doni la χ^2 -adedo mem. Ĉar ĝi estas sumo, po unu adiciato por ĉiu ĉelo, oni povas konsideri tiun adiciaton kiel indikilon por esprimi la malprobablon de la ĉelo: ju pli granda estas la indikilo, des pli granda esta la distanco inter la observita ofto kaj la ekspektita ofto; do, des pli malakceptinda estas H_0 laŭ la informo en tiu ĉelo.

Ekzemple, estu la oftotabelo

3	8	7	1
6	16	14	3
9	23	20	29

Se oni uzas komputilprogramon (ekzemple R, http://www.r-project.org) por testi nedependecon en tiu tabelo per la Pearsona χ^2- adedo, la rezulto estas P-valoro egala al 0,01577 (laŭ la asimptota distribuo de la adedo; resamplado donas valorojn ĉirkaŭ 0,01473). Se konsideri la tradician signifikancnivelon 0,05, oni malakceptu H_0. Kiuj ĉeloj kulpas pri la malakcepto? Jen la tabelo kun indikiloj $\chi^2_{ij} = (O_{ij} - E_{ij})^2 / E_{ij}$:

0,12	0,39	0,35	2,73
0,18	0,6	0,54	4,23
0,21	0,7	0,63	4,96

Klare videblas ke la ĉeloj dekstraj havas multe pli da kulpeco ol la ceteraj, kiuj proksimume estas proporciaj: en la oftotabelo, la dua linio kvazaŭ duoblas la unuan, kaj la tria kvazaŭ trioblas la unuan (krom pro la dekstra kolumno).

Je kelkaj fojoj, oni bezonas elekti klaran sojlon por justigi decidon kulpigi iujn ĉelojn kaj ne aliajn. En la ekzemplo, ĉu elekti $\chi^2_{ij} > 3$, kaj nomi "signifikancaj" la du lastajn ĉelojn? Aŭ ĉu elekti 1 kiel sojlon, kaj konsideri signifikancaj la tri ĉelojn de la lasta kolumno? Al mi okazis plurajn fojojn ke biologiistoj petis iun kriterion similan al P-valoro, por fari tiajn decidojn.

La P_{ij}-valoro

P-valoro estas la probablo, sub H_0, akiri muestron almenaŭ tiel "strangan" (laŭ H_0) kiel la observitan muestron. Se tiu probablo estas malgranda (ekzemple, sub 0,05) oni konsideru la observitan tabelon sufiĉe stranga sub H_0, do oni malakceptu H_0: ĝi ne povas esti vera, ĉar la efektive observita muestro malkongruas kun ĝi.

Ĉar jam kutimas la kalkulo, per komputila resamplado, de P-valoro de χ^2- adedo (asimptota distribuo ne taŭgas se multaj ĉeloj enhavas malgrandajn oftojn), oni povas profiti la resamplan buklon por kalkuli indikilon pri la "strangeco" au "malprobablo" de ĉiu ĉelo, supozante ke H_0 veras. Por kalkuli la P-valoron de tabelo per resamplado, oni samplas tabelojn generitajn sub H_0, kaj kun la samaj oftosumoj laŭ linioj kaj laŭ kolumnoj, kiel la observita tabelo. En R estas specifa funkcio por tia resamplo, nomata r2dtable. Por ĉiu samplita tabelo oni kalkulu la χ^2- adedon; tiel oni akiras proksimuman distribuon de la adedo, do oni povas stimi la P-valoron per la ofto de adedaj resamplaj valoroj pli grandaj ol la adeda valoro por la observita tabelo.

Simile, por ĉiu tabela ĉelo (i,j), oni apriore havas la observitan absolutan ofton O_{ij}, kaj la absolutan ofton E_{ij} ekspektitan sub H_0. Por ĉiu samplita tabelo, oni konstatu ĉu la ĉelo (i,j) enhavas ofton S_{ij} almenaŭ tiel "ekstreman" kiel O_{ij} (tio estas, ne malpli "ekstreman" ol O_{ij}). "Ekstrema" signifas "malpli probabla sub H_0 laŭ la sama direkto kiel O_{ij}"; tio estas, se $O_{ij}>E_{ij}$, oni kalkulu la relativan ofton de $S_{ij}\geq O_{ij}$; se $O_{ij}<E_{ij}$, oni kalkulu la relativan ofton de $S_{ij}\leq O_{ij}$. Tiu ofto estus la stimo de ia P-valoro por la ĉelo (i,j); mi nomos ĝin "P_{ij}-valoro".
Por la ekzempla tabelo, jen la P_{ij}-valoroj post 200.000 resamploj:

0,45970	0,28446	0,30701	0,03050
0,38986	0,17830	0,20353	0,00361
0,30412	0,08039	0,10185	0,00004

La informo estas esence la sama kiel en la antaŭa tabelo, kun la χ^2-indikiloj. Sed ŝajnas ke iuj preferas ĉi tiun esprimmanieron, kie oni povas uzi "tradiciajn" sojlojn kiel 0,05 aŭ 0,01, por decidi pri la elstaraj ĉeloj.

Komputile

Por akiri la suprajn rezultojn, oni povas tajpi en R-seanco:

```
tabelo <- rbind (c (3,  8,  7,  1),
                 c (6, 16, 14,  3),
                 c (9, 23, 20, 29))
testo <- chisq.test (tabelo)
(testo$observed – testo$expected) ^ 2 / testo$expected
hhikv.testo (tabelo, sim=T, B=200000)$p.valoroj
```

La funkcio hhikv.testo estas modifaĵo de la R-funkcio chisq.test, por kalkuli P_{ij}-valorojn. La modifitajn liniojn oni rimarkigas per grasaj literoj sur la jena printaĵo:

```
hhikv.testo <-
function (x, y = NULL, correct = TRUE, p = rep(1/length(x), length(x)),
        rescale.p = FALSE, simulate.p.value = FALSE, B = 2000)
{
 pv.matrico <- "Mi signifas nenion" # se ghi ne estas modifota, pro simulate=FALSE
 DNAME <- deparse(substitute(x))
 if (is.data.frame(x))
   x <- as.matrix(x)
 if (is.matrix(x)) {
   if (min(dim(x)) == 1)
     x <- as.vector(x)
 }
 if (!is.matrix(x) && !is.null(y)) {
   if (length(x) != length(y))
     stop("'x' and 'y' must have the same length")
   DNAME <- c(DNAME, deparse(substitute(y)))
   OK <- complete.cases(x, y)
   x <- factor(x[OK])
   y <- factor(y[OK])
   if ((nlevels(x) < 2) || (nlevels(y) < 2))
     stop("'x' and 'y' must have at least 2 levels")
   x <- table(x, y)
   names(dimnames(x)) <- DNAME
   DNAME <- paste(DNAME, collapse = " and ")
 }
 if (any(x < 0) || any(is.na(x)))
   stop("all entries of 'x' must be nonnegative and finite")
 if ((n <- sum(x)) == 0)
   stop("at least one entry of 'x' must be positive")
 if (simulate.p.value) {
   setMETH <- function() METHOD <<- paste(METHOD, "with simulated p-value\n\t (based on",
                                          B, "replicates)")
   almost.1 <- 1 - 64 * .Machine$double.eps
 }
 if (is.matrix(x)) {
   METHOD <- "Pearson's Chi-squared test"
   nr <- nrow(x)
   nc <- ncol(x)
   sr <- rowSums(x)
   sc <- colSums(x)
   E <- outer(sr, sc, "*")/n
   dimnames(E) <- dimnames(x)
   pv.matrico <- matrix (1, nrow=nr, ncol=nc)
   if (simulate.p.value && all(sr > 0) && all(sc > 0)) {
     setMETH()
     STATISTIC <- sum(sort((x - E)^2/E, decreasing = TRUE))
     PARAMETER <- NA
     hazarda <- r2dtable (B, sr, sc)
```

```
    tmp <- numeric (B)
    for (i in 1:B) {
      pv.matrico <- pv.matrico + ifelse (x > E,
                                hazarda[[i]] >= x,
                                hazarda[[i]] <= x)
      tmp[i] <- sum(sort((hazarda[[i]] - E)^2/E, decreasing = TRUE))
    }


    pv.matrico <- pv.matrico/(B+1)
    PVAL <- (1 + sum(tmp >= almost.1 * STATISTIC))/(B +
                    1)
  }
  else {
    if (simulate.p.value)
      warning("cannot compute simulated p-value with zero marginals")
    if (correct && nrow(x) == 2 && ncol(x) == 2) {
      YATES <- 0.5
      METHOD <- paste(METHOD, "with Yates' continuity correction")
    }
    else YATES <- 0
    STATISTIC <- sum((abs(x - E) - YATES)^2/E)
    PARAMETER <- (nr - 1) * (nc - 1)
    PVAL <- pchisq(STATISTIC, PARAMETER, lower.tail = FALSE)
  }
}
else {
  if (length(x) == 1)
    stop("'x' must at least have 2 elements")
  if (length(x) != length(p))
    stop("'x' and 'p' must have the same number of elements")
  if (any(p < 0))
    stop("probabilities must be non-negative.")
  if (abs(sum(p) - 1) > sqrt(.Machine$double.eps)) {
        if (rescale.p)
          p <- p/sum(p)
        else stop("probabilities must sum to 1.")
       }
  METHOD <- "Chi-squared test for given probabilities"
  E <- n * p
  names(E) <- names(x)
  STATISTIC <- sum((x - E)^2/E)
  if (simulate.p.value) {
    setMETH()
    nx <- length(x)
    sm <- matrix(sample(1:nx, B * n, TRUE, prob = p),
                 nrow = n)
    ss <- apply(sm, 2, function(x, E, k) {
      sum((table(factor(x, levels = 1:k)) - E)^2/E)
    }, E = E, k = nx)
    PARAMETER <- NA
    PVAL <- (1 + sum(ss >= almost.1 * STATISTIC))/(B +
                    1)
```

```
  }
  else {
    PARAMETER <- length(x) - 1
    PVAL <- pchisq(STATISTIC, PARAMETER, lower.tail = FALSE)
  }
 }
 names(STATISTIC) <- "X-squared"
 names(PARAMETER) <- "df"
 if (any(E < 5) && is.finite(PARAMETER))
   warning("Chi-squared approximation may be incorrect")
 structure(list(statistic = STATISTIC, parameter = PARAMETER,
              p.value = PVAL, method = METHOD, data.name = DNAME, observed = x,
              expected = E, residuals = (x - E)/sqrt(E), p.valoroj = pv.matrico),
         class = "htest")
}
```

BIBLIOGRAFIO

BAVANT, Marc. Matematika vortaro kaj oklingva leksikono. Prago: Kava-Pech, 2003.

QUEDNAU, H.D. *Enkonduko al la Deskripta Statistiko*: http://w3.forst.tu-muenchen.de/~quednau/deskripta_statistiko/

R Development Core Team. *R: a language and environment for statistical computing*. Vieno: R Foundation for Statistical Computing, 2009.

REIERSØL, Olav. *Matematika kaj stokastika terminaro esperanta*. Oslo: Institute of Mathematics, University of Oslo, 1994.

REVO. *Reta Vortaro*: http://purl.org/net/voko/revo/