Bonetti & Pagano (2005?). The interpoint distance distribution, ... disease clustering Statistics in Medicine.

Bahjat F. Qaqish
(Bahjat_Qaqish@unc.edu)

The University of North Carolina at Chapel Hill

Basic Ideas

- n points (events, cases of disease) in a region S.
- Look at the $n(n-1)/2$ distances between points
- Compare the observed distribution to what would be expected under H_0
- If H_0 rejected, try to locate the cluster(s)

The continuous case (sec 2.1)

- n points $X_1, \cdots, X_n$ iid from a continuous density
- $F(d) := \text{pr}(||X_1 - X_2|| \leq d)$, where $||.||$ is some norm.
- Borel (1925), Bartlett (1964): derived F for uniform points on the square and the circle. A lot more results from geometric probability are available.
- The empirical cumulative distribution function (ecdf):

$$F_n(d) := \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} I(||X_i - X_j|| \leq d)$$

  a V-statistic. This is for one point d.
- Being a mean, $F_n(d)$ converges to its expected value $F(d)$. The quantities averaged are correlated, but not enough to hurt the consistency.
- For a finite fixed set of points $d = (d_1, \cdots, d_m)$:

$$\sqrt{n}(F_n(d) - F(d))$$

converges in distribution to a multivariate Gaussian.

- For $d > 0$ the process

$$\sqrt{n}(F_n(d) - F(d))$$

converges weakly to a Gaussian process.

Stationary processes

Ripley's K function

$$K(d) = \frac{1}{\lambda}\mu(d)$$

where $\mu(d)$ is the expected number of events within distance d from an arbitrary event.

$\lambda :=$ the intensity $=$ the expected number of events per unit area

$K(d)$ can be estimated for $d \leq$ half the maximum distance inside S.

The discrete case (sec 2.2)

Locations $l_1, \cdots, l_k$ (non-random)

Distances $d_{ij} := ||l_i - l_j||$

Probabilities $p_1, \cdots, p_k$, sum to 1

Numbers of events $(n_1, \cdots, n_k)$ is multinomial with parameters (n, p)

D is the random distance between two independent events

D is discrete, and

$$\text{pr}(D = d_{ij}) = p_i p_j$$

The cumulative distribution function:

$$F_n(d; p) := \sum_{i=1}^{k} \sum_{j=1}^{k} p_i p_j I(d_{ij} \leq d)$$

The empirical cumulative distribution function (ecdf):

$$F_n(d; \hat{p})$$

where $\hat{p}_i := \frac{n_i}{n}$.

Tests (sec 2.3)

H_0 specifies p, and hence $F(d; p)$.

Usually p is the observed distribution in *non-cases*, usually census data [could be model-based]

Fix a grid of points $d := (d_1, \cdots, d_m)$.

Residuals (vector): $e := F_n(d) - F(d)$

Quadratic form: $\tilde{M} = e^\top \Sigma^- e$

$\approx \chi^2_r$ for large n.
r is the rank of Σ.

Claim: $M = e^\top S^- e$ works better (in terms of the χ^2 approx.). No justification given

S is the sample covariance of e from 1000 resamples of size n with replacement [bootstrap?] [Resampling is not under H_0]

f_n is a numerical derivative of F_n, an estimate of the density.

F_n ia AN, so is f_n (linear combinations of F_n).

The distance $||f_n - f||$:
L_2 norm
χ^2
KL (E[log liklihood ratio])

Disease clustering (sec 3.1)

Figure 1, F_n for
1980 census ($n = 10^6$) and
1978-1982 cases ($n = 581$)

- Centers $1, \cdots, k$, and p_i = the population proportion
  Sum $(O - E)^2/E$
  Pearson's X^2
  Geography plays no role
- Tango (1995)
  y_i/n_i = #cases / #subjects in region i

$$e_i := \frac{y_i}{y_.} - \frac{n_i}{n_.},$$

$$T = e^\top W e$$

a QF with respect to weights that depend on distance.

W is a weight matrix with $w_{ii} = 1$ and $w_{ij} = \exp(-d_{ij}/\alpha)$.

- Whittemore et al.(1987)
  δ = the mean distance between pairs of distinct points = the mean of $n(n-1)/2$ distances. Compare to the mean and variance under H_0.

Upstate NY Data (sec 3.2)

Figure 2, f_n

f_n has higher peaks than f at 0, 60 and 110 km.

p-values:
$T : p = 0.000$
$DC : p = .944$
$\delta : p = 0.804$

Locating clusters

Score(i) defined for the ith location, such that it sums to M.

A lot of arbitrariness, but seems to detect a couple of clusters detected by other methods.

But Score(i) still involves a sum over all locations!

Comments

Moving all cases by rigid rotations, reflections and translation deos not affect M.

Thus M detects whether the clustering of cases is similar to clustering in the general population, not that there is a clustering of cases in the sense of Kulldorf (areas of high incidence).

A lot of room for new ideas!