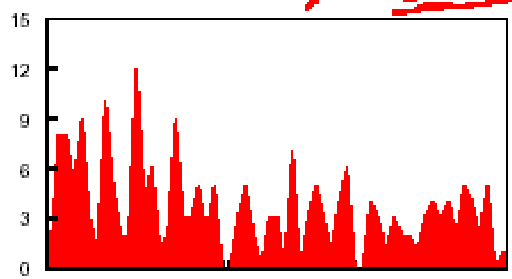
Análisis espacial de epidemias: Patrones aleatorio (para datos discretos)

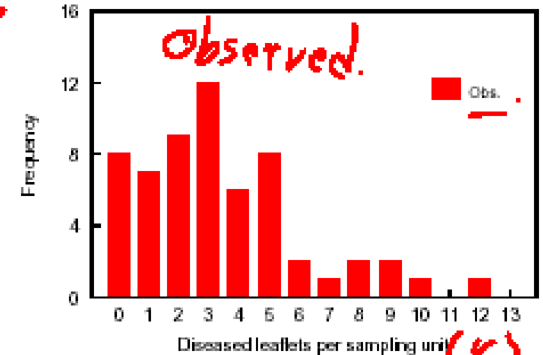
- La probabilidad de que una planta esté enferma es independiente del estatus de enfermedad de las otras plantas
  - **Esto es, *p* es constante**
- El estatus de una planta no está relacionada con es estatus de planta enferma de sus vecinas
- Conociendo el estatus de una planta enferma no provee información sobre el estatus de las otras plantas
- Si p es constante, los individuos enfermos por unidad de muestreo (*Y*) tienen una distribución binomial
- **Distribución** o probabilidad de distribución:
  - **(Para variables discretas aleatorias) una fórmula matemática da la probabilidad de cada valor de la variable**
- Las distribuciones son evaluadas por comparación de las frecuencias observadas (*O*) de individuos enfermos para predecir la frecuencia esperada; *E*)

Phomopsis leaf blight of strawberry (Turechek & Madden, 1999, and other papers). - Transect through field. N=59, n=15; $\hat{p}$= **0.226** (=200/[15*59]) $= \bar{y} = \frac{\bar{Y}}{15} = \frac{3.39}{15}$

$$s_y^2 = \frac{\sum (y_i - \bar{y})^2}{N-1} = 0.0325$$

Y

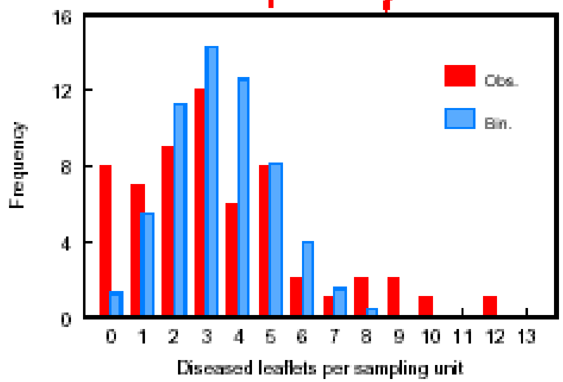




(Y)

1	8	8	5	9	3	1	10	5	2
2	12	4	6	1	2	9	3	3	5
2	5	0	0	3	5	2	0	3	3
0	7	0	3	5	3	1	4	6	0
0	4	3	1	3	2	2	1	3	4
3	4	2	5	4	2	5	0	1	

$$Pr(y) = \binom{n}{y} p^y (1-p)^{n-y}$$



$$p = 0.226$$

$$Pr(0) = \binom{15}{0} (.226)^0 (1-.226)^{15}$$

$$= 1 \cdot 1 \cdot .774^{15} = .02143$$

$$E(0) = Pr(0) \cdot N = .2143 \cdot 59 = 1.26$$

$$Pr(1) = \binom{15}{1} .226^1 \cdot .774^{14}$$

$$= 15 \cdot .226 \cdot .0277$$

$$= .0939$$

$$E(1) = .0939 \cdot 59$$

$$= 5.5$$

Distribución binomial:

- Media estimada *Y:np*
- Varianza estimada de a proporción

  $p(1-p)/n = s_{bin}^2$
- Test de bondad del ajuste
  - Se usa el χ^2
  - Un buen ajuste= valor chico de χ^2
    - Df=basado en el número de clases
    - De Y (no basado en N)
  - Por ejemplo, χ^2 = 15.9 (df=4; P<0.01)
    - Pobre ajuste: - muy alto en el medio

      - muy bajo cerca de 0
  - Por lo tanto, alguna evidencia de no aleatoriedad

$\underline{Bin}$: Y $\quad Var(Y) = np(1-p)$

$Var(Y) = np(1-p)$

if Y has $\underline{Bin}$. , then

variance of $\bar{y} = \frac{Y}{n}$

$Var(\bar{y}) = \frac{var(Y)}{n^2}$

$var(\bar{y}) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$

$$se(\hat{p}) = se(\bar{y}) = \sqrt{\frac{var(y)}{N}}$$

$$= \sqrt{\frac{\bar{y}(1-\bar{y})}{N}}$$

binomial

y, Y

Análisis espacial de epidemias: Patrones no aleatorios (para datos discretos)

- La probabilidad de que una planta esté enferma **no** es independiente del estatus de enfermedad de las otras plantas
  - **Esto es, *p* NO es constante, pero es una variable constante**
- El estatus de una planta está relacionada con es estatus de planta enferma de sus vecinas (la correlación del estatus de no enferma de una planta dada respecto a su vecina es cero)
  - Si una planta dada está enferma, hay una tendencia de que las plantas vecinas estén enfermas
- Conociendo el estatus de una planta enferma provee alguna información sobre el estatus de las otras plantas
- Aproximación: hay que especificar una distribución estadística para *p* (dado que ahora es una variable aleatoria)
  - Si *p* tiene una distribución **beta,** luego *Y* tiene una **distribución beta-binomial**
  - Esta distribución tiene una fórmula bastante complicada

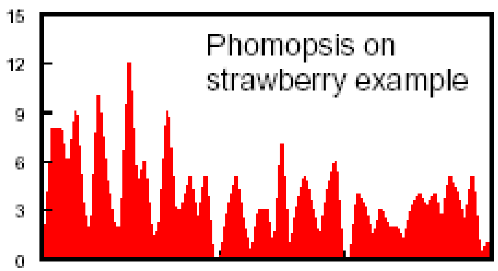
Distribución beta-binomial (dos parámetros)

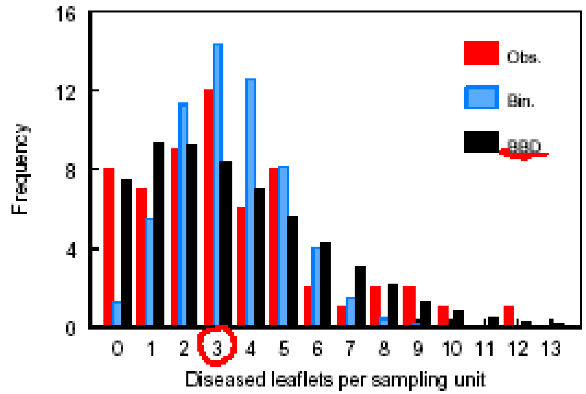
- *p* (probabilidad media de que una planta esté enferma)
- θ(parámetro que indica heterogeneidad o agregación)
  - θ =0 (se reduce a la binomial)
  - θ>0 (agregado, cluster)
  - Puede ir a infinito pero 0.2 es grande

$0 \rightarrow \infty$

Expected $\hat{p} = 0{,}226$ $\hat{\theta} = 0.14$

Y





- Chi-cuadrado es el test para bondad de ajuste: 6.2 (no significativo)
  - No significativo resulta en un buen ajuste a la distribución beta-binomial (agregado)
  - Pero un ajuste significativo resulta en una binomial lo que sugiere datos agregados

Ajustando datos a una distribución discreta

- Máxima verosimilitud es generalmente lo mejor, pero esto puede requerir programas de computación muy especializados (un método alternativo es el beta-binomial)
- Métodos más simples funcionan mejor para muchos propósitos
  - El método del momento
  - Algunas veces, los parámetros estimados son muy cercanos por diferentes métodos
  - Los métodos del momento no pueden más que estimar la media y la variancia de una muestra

$$\frac{\bar{y}(1-\bar{y})}{n} = s^2_{bin}$$

Binomial

$$\boxed{\hat{p} = \bar{y} = \frac{\bar{Y}}{n}}$$

Beta-binomial

$$\hat{p} = \bar{y} = \frac{\bar{Y}}{n}$$

$$\hat{\theta} = \frac{s_y^2 - \bar{y}(1-\bar{y})/n}{\bar{y}(1-\bar{y}) - s_y^2} = \frac{s_Y^2 - n\bar{y}(1-\bar{y})}{n^2\bar{y}(1-\bar{y}) - s_Y^2}$$

- Aunque es útil ajustar directamente distribuciones a datos, y determinar su bondad de ajuste, esa aproximación no es necesaria
- En particular, uno puede utilizar propiedades de la distribución beta-binomial para probar agregación y cuantificar el grado de agregación
- Por ejemplo, es muy importante considerar la estimación de la varianza de los datos discretos s_y^2 o s_Y^2

$$s_y^{\ 2} = \frac{\sum\left(Y_i - \bar{Y}\right)^2}{N-1} \qquad\qquad s_Y^{\ 2} = \frac{\sum\left(Y_i - \bar{Y}\right)^2}{N-1}$$

- La variancia de una variable con la distribución beta-binomial es:

$$s^2_{\beta bin} = \left[\frac{\hat{p}(1-\hat{p})}{n}\right]\left(\frac{1+n\hat{\theta}}{1+\hat{\theta}}\right) = \left[\frac{\bar{y}(1-\bar{y})}{n}\right]\left(\frac{1+n\hat{\theta}}{1+\hat{\theta}}\right)$$

Note: $\frac{\hat{p}(1-\hat{p})}{n} = \frac{\bar{y}(1-\bar{y})}{n} = s^2_{bin}$ variance of binomial variable

$$s^2_{\hat{p}\,\text{bin}\,Y} = \underbrace{\left[n\hat{p}(1-\hat{p})\right]}_{\text{bin}(Y)} \cdot \underbrace{\left(\frac{1+n\theta}{1+\theta}\right)}_{\text{scaling factor}}$$

$$s_Y^2 = \left(s_{bin}^2 = \boxed{\frac{\hat{p}(1-\hat{p})}{n}} \left(\frac{1+n\hat{\theta}}{1+\hat{\theta}}\right) = \left[\frac{\bar{Y}(1-\bar{Y})}{n}\right] \left(\frac{1+n\hat{\theta}}{1+\hat{\theta}}\right) \right)$$

Note: $\frac{\hat{p}(1-\hat{p})}{n} = \frac{\bar{Y}(1-\bar{Y})}{n} = s_{bin}^2$ variance of binomial variable

- Por lo tanto la variancia de la beta-binomial iguala a la variancia binomial (para un patrón aleatorio) por un factor de escala que se eleva con el incremento de la agregación
- A θ =0, la variancia beta-binomial iguala a la variancia de la binomial
- A θ >0 la variancia de la beta-binomial es más grande que la variancia binomial
- Dado que el factor de escala puede tomar algún valor (en principio) la variancia binomial puede ser igual al valor real de una muestra s_y^2
  - De hecho, modificando la fórmula de abajo, puede ser usada para estimar el momento de Θ

- Un estadístico muy útil para caracterizar agregación es la relación de la variancia observada (la cual no está basada en supuestos acerca de la distribución) y la variancia estimada para una variable con una distribución binomial (ej. Para una situación aleatoria)

$$D = \frac{s_y^2}{\bar{y}(1-\bar{y})/n} = \frac{s_y^2}{\hat{p}(1-\hat{p})/n} = \frac{s_y^2}{s_{bin}^2}$$

$$\boxed{s_y^2 = \frac{\sum (y_i - \bar{y})^2}{N-1}}$$

$$s_{bin}^2 = \frac{\bar{y}(1-\bar{y})}{n} = \frac{\hat{p}(1-\hat{p})}{n}$$

- D es conocido como el índice de dispersión
- Recordar que cualquier variancia dada puede ser escrita como un producto de una variancia binomial y un factor de escala, por lo tanto D también es igual:

$$D = \boxed{\frac{s^2_{pbn}}{s^2_{bn}}} = \frac{\cancel{\left(\frac{\bar{y}(1-\bar{y})}{n}\right)}\left(\frac{1+n\theta}{1+\theta}\right)}{\cancel{\left(\frac{\bar{y}(1-\bar{y})}{n}\right)}}$$

$$\boxed{D = \frac{1+n\theta}{1+\theta}}$$

- **Una evaluación muy simple de agregación**
  - Conseguir D (de una variancia observada y una variancia binomial)
    - D=1: aleatorio
    - D>1: agregada
    - D<1: regular (uniforme)

      -El mínimo de D es 0

$\max D = n$

$$s_y^2 = \frac{\sum (y-\bar{y})^2}{N-1}$$

$$D = \frac{s_y^2}{\bar{y}(1-\bar{y})/n} = \frac{s_y^2}{s_{bin}^2}$$

  - Test de agregación
    - **(N-1)D**

      - Chi cuadrado tiene una distribución con grados de libertad de N-1 (si es binomial)

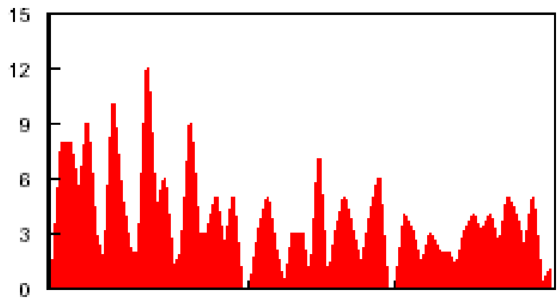
      - H_0: aleatoria (binomial)

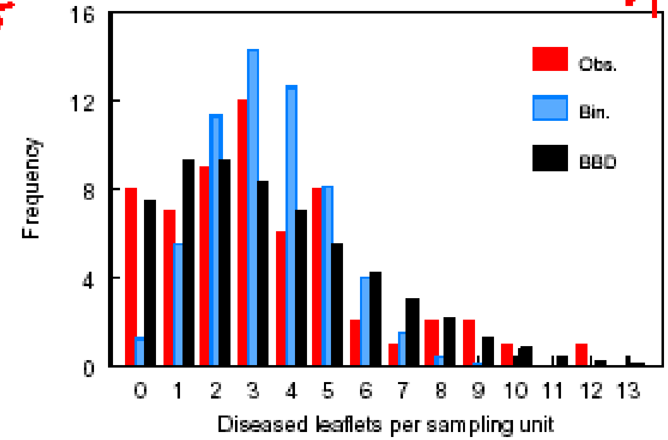
      - Si (N-1)D>valor crítico de chi cuadrado, luego se concluye que es agregado

$$\boxed{\begin{aligned}\bar{Y} &= \vec{p} = 0.226 \\ \vec{\theta} &= 0.14\end{aligned}}$$

$s_y^2 = 0.0325$ (square of standard deviation)

$$s^2_{bin} = \frac{.226(1-.226)}{15} = 0.01166 = \frac{\vec{p}(1-\vec{p})}{n}$$





$$D = \frac{.0325}{.01166} = 2.78$$

$$(N-1)\cdot D = (59-1)\cdot 2.78 = 161$$

$[\chi^2_{58} = 76.8]$ Significant (>1)

P = .05

Significado de la agregación basado en la varianza y/o distribuciones discretas

- La distribución binomial no es la única distribución que puede describir datos de incidencia de enfermedad agregada, pero es la más común
  - – Tiene propiedades teóricas muy útiles (no discutidas aquí)
  - – Un caso especial (θ =0) es la binomial (aleatorio)
- Distribuciones de este tipo caracteriza explícitamente heterogeneidad de la variable aleatoria
  - – Cuando θ >0, la variable es´tá sobre diseminada (por lo tanto θ es una medida de sobre dispersión o grado de heterogeneidad
- La beta-binomial (o similar) caracteriza el patrón de enfermedad a una escala espacial de la unidad de muestreo o más pequeña
  - – Esto es θ representa agregación o individuos enfermos con unidades de muestreo no a través de la unidad de muestreo
  - – **PATRONES DE PEQUEÑA ESCALA** (ej. pequeñas manchas)
  - – Si los datos fueron mapeados (no requerido porque todo esto funciona para datos colectados de muestra laxas), uno puede ver necesariamente grandes manchas de alta enfermedad y baches de baja enfermedad
- **Mayor agregación dentro de unidades de muestreo se manifiesta por la gran variabilidad entre unidades de muestreo**

Significado de la agregación basado en la variancia y/o distribuciones discretas

- Patrones de escala chicos pueden ser claros considerando **la correlación intra-cluster (*p*)**
  - La correlación del estado de enfermedad de individuos dentro una unidad de muestreo (un promedio a través de unidades de muestreo)
  - Tendencia de individuos dentro de una unidad de muestreo a tener el mismo valor
- Esto se puede ver: ***p* = *θ*/(1+ *θ*)**
  - ***p* = 0:** no hay correlación (no agregado)
  - ***p* > 0:** Agregación (máximo de 1)

$$D = \frac{1+n\theta}{1+\theta}$$

Por lo tanto

$$s_y^2 = \left[\left(\frac{\bar{p}(1-\bar{p})}{n}\right)\right]\left[\left(1+\rho(n-1)\right)\right] \quad \text{or} \quad D = \frac{s_y^2}{s_{bin}^2} = \left(1+\rho(n-1)\right)$$

$$= \left(s_{bin}^2\right)\left(1+\rho(n-1)\right) \qquad \rho = 1$$

$\rho = .14/1.14 = 0.115$

example

Resumen

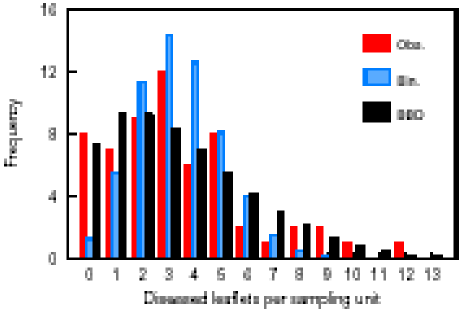
- Existen múltiples formas de decir casi la misma cosa acerca de un set de datos en términos de hetrogeneidad / sobreposición - patrones en pequeña escala
  - Algunas veces solo una cuestión de preferencia
- Un ajuste completo a un modelo de distribución de datos es más informativo (más información que solo medias y varianzas), pero es más desafiante
  - La bondad del ajuste no puede ser siempre determinada
- Nota: Hay mucho más de patrón que agregación a pequeña escala
  - Ejemplo, grandes manchas que se extiende sobre unidades de muestreo múltiples
  - O mezclas de manchas grandes y pequeñas

$\hat{p} = .226$

$\hat{\theta} = .14$

$\hat{\rho} = .115$

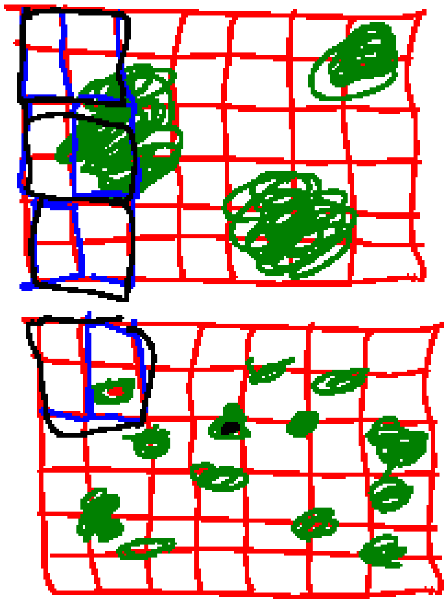
$D = 2.78$



Patrones de enfermedad

- Análisis basados en datos de mapeos intensivos
  - Considera solo muestreo en cluster (n individuos por unidades de muestra) N unidades de muestreo – incidencia (Y_i/N)
- De nuevo, relacionado con arreglos de individuos enfermos
- Nota: cuando la probabilidad de que una planta esté enferma no es constante (por lo que el estatus no está relacionado con el estatus de otros), ocurren patrones no aleatorios
- Primariamente, el interés es encontrar agregación (clusters, manchas, ..)
  - Los métodos previos han caracterizado patrones a pequeñas escalas (representados por correlación intra cluster e índices relacionados)
  - Los métodos previos no proveen información sobre grandes escalas espaciales (grandes áreas)
    - » De hecho cualquier arreglo de los N conteos dan el mismo D, p, etc
  - **Ahora, nuestro interés está en: Tendencia para observaciones desde localidades cercanas para tener magnitudes similares comparadas con localidades alejadas**

- Algunos métodos más viejos tratan combinando unidades de muestreo dentro de grupos, y determinando como índices (D, etc,) cambian con el tamaño de la unidad de muestreo (solamente como parcialmente informativo)
- Con el avance de la geoestadística y otros análisis espaciales, tales aproximaciones no son necesarias ahora
- Para el análisis típico, Y (o y) pueden ser tanto una variable discreta (conteo: no binaria) o variable continua aleatoria
  - Incidencia, densidad o severidad
  - Clave: referencia espacial de las unidades de muestreo
- Dos métodos mayores
  - **Análisis de la autocorrelación espacial**
  - **Análisis de Semivariogramas**
- Por supuesto, aquí hay otras alternativas para las unidades de muestreo consistente en individuos solos



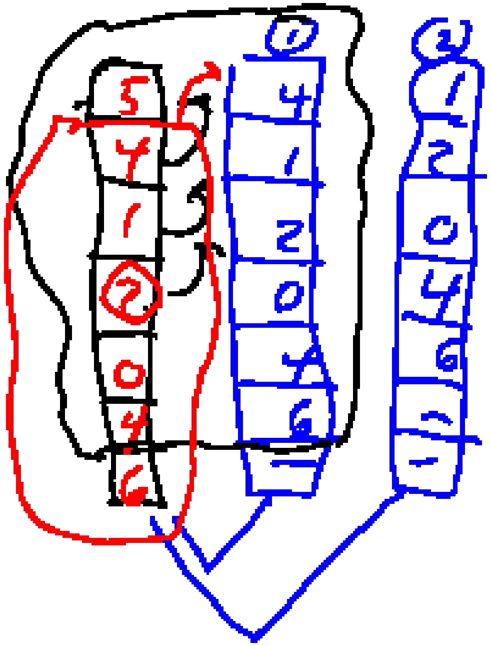
Análisis espacial de epidemias: Patrones de enfermedad

- La aproximación estadística usada para análisis de patrones (porque entre otras cosas, las fuentes claramerente identificables de inóculo no son definibles)
- Para datos de incidencia (discreto, ej. Conteo con límite superiro [n]) uno puede utilizar propiedades de distribuciones estadísticas para cuantificar los patrones
  - En general, el análisis caracteriza atributos de patrones a pequeña escala (grado de agregación de individuos enfermos dentro de unidades de muestreo) --- ***D*, *θ*, *p***
- Con unidades de muestreo de mapeo intensivo, hay muchos métodos estadísticos posibles para ser usados
  - En general , el análisis caracteriza patrones de gran escala (grados de similaridad (o disimilaridad) de valores en unidades de muestreo)
  - Las variables analizadas aquí pueden ser discretas o continuas

Autocorrelación espacial

- Mide el grado de asociación en Y entre unidades de muestreos vecinos
- Para vecinos inmediatos (cada unidad de muestreo con los siguientes a él), ***r(1)***
  - Determinado más comúnmente determinado
- Para el próximo vecino más inmediato , ***r(2)***
  - Así sucesivamente
- Mientras más grande ***r(.)***, más agregación

$$\hat{r}_{(i)} = \frac{\dfrac{\sum (Y_i - \bar{Y})(Y_{i+1} - \bar{Y})}{N_1}}{\dfrac{\sum (Y_i - \bar{Y})^2}{N}}$$



N_i=Número de pares para vecinos inmediatos

Autocorrelación espacial general

$$\hat{r}_{(i)} = \frac{\dfrac{\sum \left(Y_i - \bar{Y}\right)\left(Y_{i+1} - \bar{Y}\right)}{N_1}}{\dfrac{\sum \left(Y_i - \bar{Y}\right)^2}{N}}$$

j = "lag" espacial

j=1 vecinos más cercanos

j=2 una unidad más lejos

Nj= número de pares de unidades

$$= \frac{Co\,\mathrm{var}\,iancia\left(Y_i, Y_{i+j}\right)}{Variancia\left(Y_i\right)} = \frac{\hat{C}(v)}{\hat{C}(0)} = \frac{\hat{C}(j)}{S_j^{\,2}}$$

Autocorrelación espacial

- Se evalúa la magnitud de ***r(1)***
  - Valores grandes indican agregación, a una escala espacial más grande que el tamaño de la unidad de muestreo
    - » El error estándar de *r(1):* $\sim(1/N_1)^{½}$
- Los resultados combinados para ***r(1), r(2), r(3), etc,*** puede distinguir entre manchas grandes y pequeñas y otros arreglos
  - Error estandar: $\sim(1/N_1)^{½}$

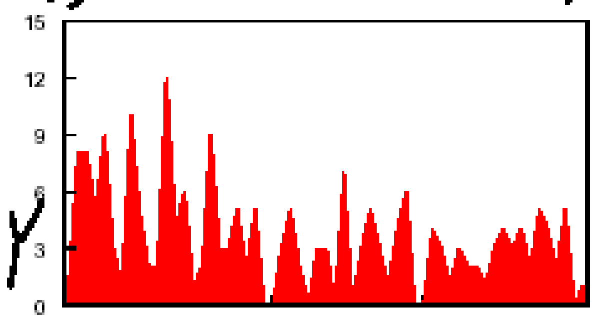
- Nota: El análisis aquí son para diferente escala distinta al análisis de heterogeneidad (beta-binomial, D, etc) que se presentó antes
- Por lo tanto los métodos pueden dar diferente resultados, dado a que ello no están caracterizando lo mismo

Time Series Analysis :

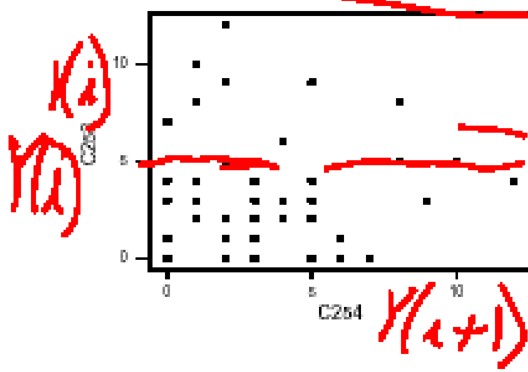
First order correlation

$r(1): -1 \leftrightarrow +1$

0 : random



$\hat{r}(1) = -.02$



$se(\hat{r}(1)) = \sqrt{\frac{1}{58}} = 0.131$

$z = \frac{-.02}{.131} = .153$

< 2

Spatial autocorrelations

ACF

```
ACF of C260

          -1.0 -0.8 -0.6 -0.4 -0.2  0.0  0.2  0.4  0.6  0.8  1.0
            +----+----+----+----+----+----+----+----+----+----+
  1  -0.020                          XX
  2  -0.087                         IXX
  3   0.251                            XXXXIXX
  4   0.148                            XXXXI
  5   0.149                            XXXXI
  6   0.080                            XXX
  7   0.035                            XX
  8   0.053                            XX
  9   0.196                            XXXXIX
 10   0.142                            XXXXI
 11  -0.206                       IXXIXX
 12   0.111                            XXXX
 13   0.045                            XX
```

r(1)

r(2)

Análisis de patrones espaciales

- **<u>Notas:</u>**
- El análisis de semivariogramas es muy muchas disciplinas
  - Semivariancia (1/2) de la variancia) de una diferencia de valores
    - » Pequeños valores indican agregación
    - » Grandes valores (=variancia) indica aleatoriedad
  - Bajo muchas condiciones, autocorrelación y semivariancias son equivalentes

$$\hat{\gamma}(j) = \frac{\sum (y_i - y_{i+j})^2}{2 N_j}$$

1, 2,

semivariances are equivalent.

$$\hat{\gamma}(j) = \hat{C}(0) - \hat{C}(j) \rightarrow \hat{\gamma}(j) = \hat{C}(0)\,(1 - \hat{r}(j)) = s_y^2\,(1 - \hat{r}(j))$$

variance

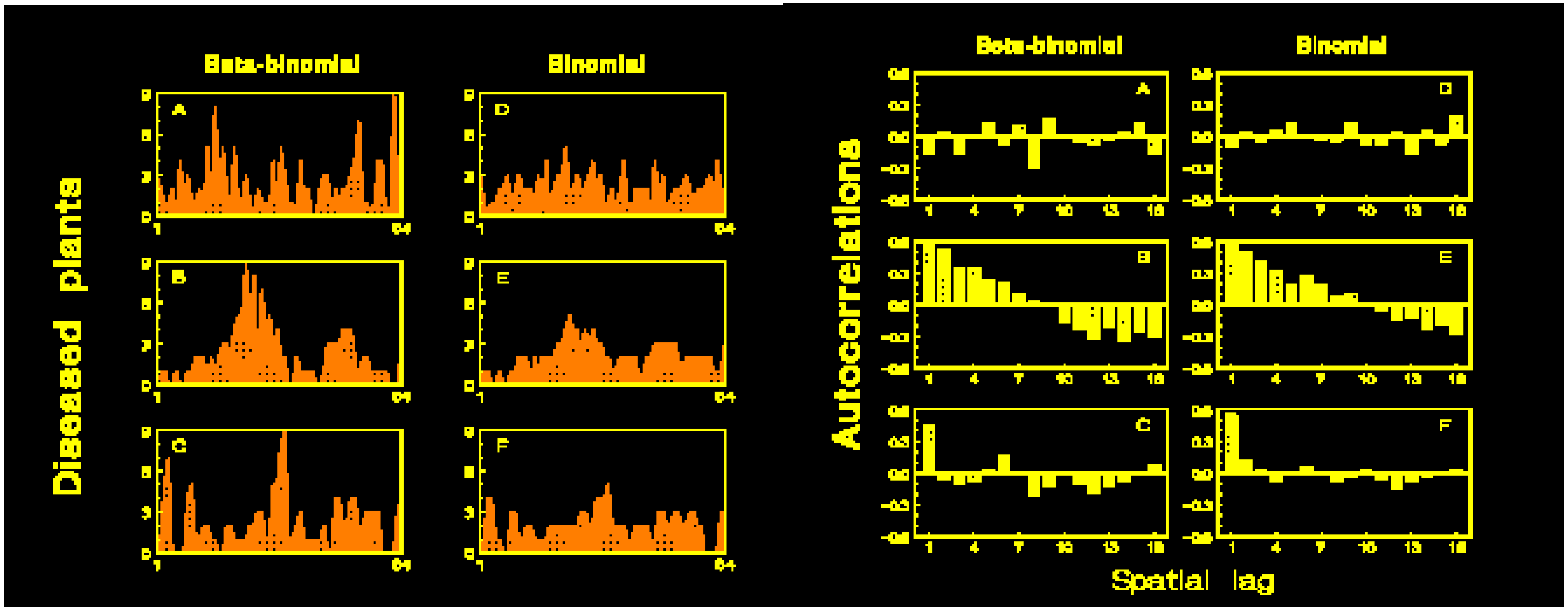
cov.

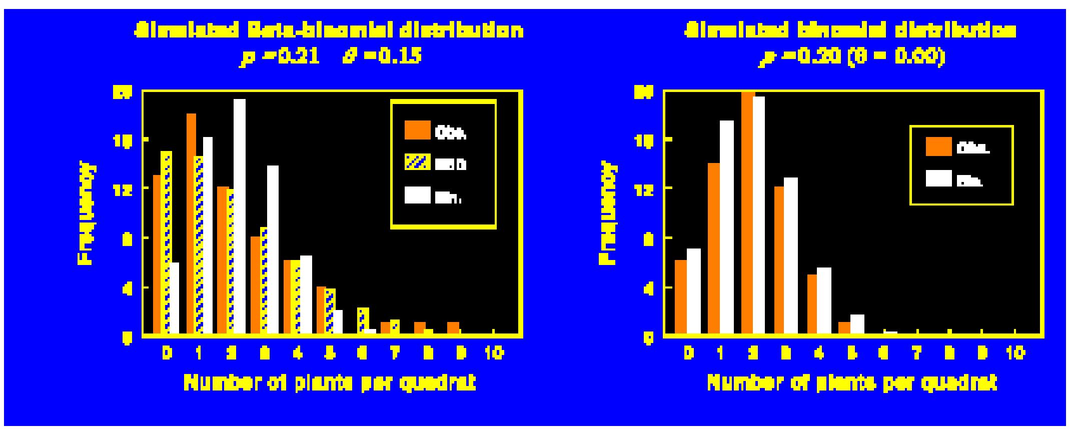
y

Resumen

- **Existen muchas maneras de caracterizar los patrones especiales de los organismos, incluyendo individuos enfermos**
- Algunas aproximaciones son dependientes del tipo de variable aleatoria (solo para datos discretos)
- Otras aproximaciones dependen de un mapeo intensivo (y no solo de muestreo laxo)
- Aquellos que trabajan con mapeo intensivo dependen de datos discretos, y pueden ser usados para mapeo intensivo
  - Caracterizar heterogeneidad (patrón a una escala de unidades de muestreo y menores)
- Los que trabajan con mapeo intensivo son aplicables a datos discretos y continuos, pero no son usados en general para muestreos laxos
  - Caracterizar patrones a la escala de unidades de muestreo y mayores
- Recordar: nosotros solamente discutimos métodos para muestreo laxo y mapeo intensivo cuando hay conteo (Y de los n) en cada unidad de muestreo
- Al menos que el muestreo sea verdaderamente aleatorio, hay una escala para el patrón
  - Esto es, podrían haber pequeños agrupamientos (aun no visible (en un sentido)) obtenidos de un mapa, pero cuantificable usando análisis de datos discretos
  - Podrían haber manchas grandes y muy grandes dispersos sobre áreas de interés, cuantificable a través de autocorrelación
- Resultados de diferentes clases de análisis pueden ser complementarios y <u>no</u> contradictorios

Métodos	Muestreo laxo	Mapeo intensivo	Discreto	Continua
Ajuste de distribuciones discretas	★	★	★	-
p	★	★	★	
θ	★	★	★	-
D	★	★	★	-
Autocorrelación	-	★	★	★
Semivariograma	-	★	★	★


Beta-binomial
Binomial
Diseased plants
Autocorrelations
Spatial lag


Simulated Beta-binomial distribution
p =0.21 θ =0.15
Frequency
Number of plants per quadrat
Simulated binomial distribution
p =0.20 (θ = 0.00)
Frequency
Number of plants per quadrat