# Categorising Ecological Community Count Data

Daniel Fernández

Shirley Pledger

Victoria University of Wellington

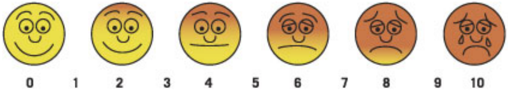
The International Biometric Society Australasian Region Conference 2015
Hobart, Tasmania
Nov. 30th - Dec. 3rd, 2015

# Outline

1. Background.
   - Count data. Variance-mean ratio.
   - Approaches: Poisson, Negative Binomial.
   - Ordinal stereotype model.
2. Advantages of categorising count data into ordinal data.
3. Categorising count data: methodology.
4. Application
   - Spider data (Van der Aart & Smeenk-Enserink, 1974).
5. Summary.

# 1. Count and Ordinal data

- Count data:
  - Count the number times an event occurs, e.g. # particular species at a certain site.
  - Non-negative integers and zero being included or not (depending on whether it is ecologically important).
  - Counts may have **no upper bound**, or have a known maximum.
- Ordinal data:
  - Answers on ordinal variable describing inherent order.
  - The order in the response categories matters.
  - For example, Braun-Blanquet scale is very common in vegetation science or Likert scale in surveys.



# 1. Count data. Variance-mean ratio

- **Variance-mean ratio** $\left(\frac{\textbf{Var}}{\textbf{Mean}}, \textbf{VMR}\right)$. Stochastic scheme for classifying count data (Rogers, 1974, ch. 1)[1]:
  - VMR>1 (variance > mean) $\Rightarrow$ *clustered* point pattern.
  - VMR=1 (variance=mean) $\Rightarrow$ dispersion follows a *random* point pattern.
  - VMR<1 (variance < mean) $\Rightarrow$ *regular* point pattern.

a. Clustered b. Random c. Regular

1. Rogers, A. Statistical Analysis of Spatial Dispersion: The Quadrat Method. Monographs in Spatial and Environmental Systems Analysis. Pion, 1974.

# 1. Count data. Variance-mean ratio.

a. Clustered b. Random c. Regular

- *Clustered* point pattern ⇒ prob. object being in quadrat **linearly** related to # objects already there.
  e.g. shoal of sardines ⇒ **negative binomial distribution**.

- *Random* point pattern ⇒ prob. object being in quadrat **independent** of the # objects already there.
  e.g. plants with well-dispersed seeds ⇒ **Poisson distribution**

- *Regular* point pattern ⇒ prob. object being in quadrat **decreases linearly** with the # objects there.
  e.g. gannet nests in a colony ⇒ **binomial distribution**.

# 1. Count data. Variance-mean ratio

a. Clustered b. Random c. Regular

- Variance-mean relationship is a critical property of count data.
- Trends in location (mean abundance) may be confounded with changes in dispersion (Warton *et al.*, 2012)[2]
- One **alternative** to deal with VMR problem $\Rightarrow$ turn count data into ordinal.

2. Warton, D. I., Wright, S. T., and Wang, Y. Distance-based multivariate analyses confound location and dispersion effects. Methods in Ecology and Evolution, 3(1):89-101, 2012.

# 2. Advantages of categorising count data

- **Possible drawbacks** what could arise from using count data.
  1. Highly sensitive to **outliers** $\Rightarrow$ negative binomial.
  2. Structurally **exclude zero counts** (e.g. hospital length of stay (in days)) $\Rightarrow$ zero-truncated models.
  3. **Excess of zero counts** $\Rightarrow$ hurdle models, zero-inflated models.
  4. One data set with **different levels of VMR** $\Rightarrow$ apply different count data models.

- **Advantages of categorising** count data into ordinal categories:
  1. Less sensitive to <u>outliers</u>.
  2. No affected by the <u>omission of zeros</u> in the data.
  3. <u>Excess of zero counts</u> $\Rightarrow$ Cumulative link random effects models $\Rightarrow$ more parsimonious (Agresti, 2010).
  4. Use of the same approach for <u>different levels of VMR</u>.

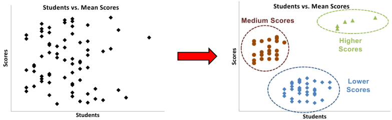
# 1. Approaches

- Data represented as a matrix $Y$ with dimensions $n \times p$ ($n$ could be sites, $p$ could be spp.)

# 1. Approaches

- Data represented as a matrix $Y$ with dimensions $n \times p$ ($n$ could be sites, $p$ could be spp.)

- Single-mode **clustering** and biclustering $\Rightarrow$ **Finite mixture models**.

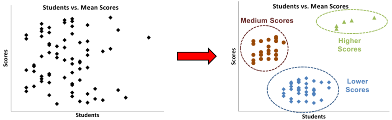



- Missing information: row/col membership $\Rightarrow$ EM algor., RJMCMC

# 1. Approaches

- Data represented as a matrix $Y$ with dimensions $n \times p$ ($n$ could be sites, $p$ could be spp.)

- Single-mode **clustering** and biclustering $\Rightarrow$ **Finite mixture models**.




- Missing information: row/col membership $\Rightarrow$ EM algor., RJMCMC
- Count data sets:
  - Assumption of Poisson distribution (Pledger and Arnold, 2014)[3]
  - Negative binomial distribution when overdispersion.

- Ordinal data sets (after categorising):
  - Assumption of ordinal stereotype model (Fernández *et al.*, 2014)[4]

3. Pledger, S. and Arnold, R. Multivariate methods using mixtures: Correspondence analysis, scaling and pattern-detection. Computational Stat. and Data Analysis, 2014.

4. Fernández, D., Arnold, R., and Pledger, S. Mixture-based clustering for the ordered stereotype model. Computational Stat. and Data Analysis, 2014.

# 3. Methodology. How Many Ordinal Categories?

Several ways of categorising:

- Simplest case: Using count data as ordinal categories (e.g. $(0, 1, 2, 3) \Rightarrow \{0, 1, 2, 3\}$).

- Large counts. Use top-coded data (e.g. $\{0, 1, \mathbf{2+}\}$). $\{0, 1+\}$: presence-absence, extreme case.

- Equally spaced cut points (e.g. $0 - 4, 5 - 9, \ldots$ or $0, 1 - 9, 10 - 99, \ldots$ with logarithmic scale).

- Replace count data by their ranks and cutting them into groups based on **percentiles**.
  - Percentiles are not strongly influenced by extreme values
  - Can be calculated even if the counts are skewed.

# 3. Methodology. Categorising Based on Percentiles

Given count data $\{y_{ij}\}$ ($i = 1, \ldots, n$ and $j = 1, \ldots, p$).

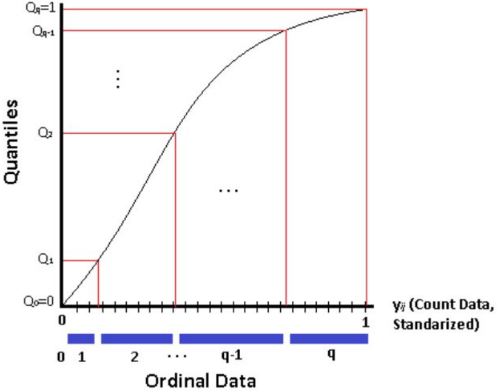
1. **Rescale** each observation $y_{ij}$, so $y_{ij}^{\mathrm{st}} \in [0, 1]$.
2. **Divide** vector $\{y_{ij}^{\mathrm{st}}\}$ into $\ell + 1$ quantiles: $Q^{(0)}, \ldots, Q^{(\ell)}$.
3. **Recode** each observation $y_{ij}^{\mathrm{st}}$ as:

$$y'_{ij} = \begin{cases} 0 & \text{if } y_{ij}^{\mathrm{st}} \in [Q^{(0)}, Q^{(1)}], \\ k & \text{if } y_{ij}^{\mathrm{st}} \in (Q^{(k)}, Q^{(k+1)}], \end{cases}$$

   where $(Q^{(k-1)}, Q^{(k)}]$ is the interval of values from vector $y_{ij}^{\mathrm{st}}$ between the $(k-1)^{\mathrm{th}}$ and $k^{\mathrm{th}}$ quantiles, for $k = 1, \ldots, \ell$.
   Each interval contains $\frac{100}{\ell}\%$ of the non-zero data.
4. **Fit** our ordinal mixture methodology to $\mathbf{Y'}$.

# 3. Methodology. Categorising Based on Percentiles

Given count data $\{y_{ij}\}$ ($i = 1, \ldots, n$ and $j = 1, \ldots, p$).



# 4. Application. Spider Dataset


*Pardosa monticola* - pin-stripe wolf spider, and it inhabits sand dunes in the Netherlands.

# 4. Application. Spider Dataset

- "Spider" abundance data (Van der Aart & Smeenk, 1974).
- 12 spider species , 28 sites .
- Original data: Count data (species abundance at site).
- Ordinal data: 4 categories.

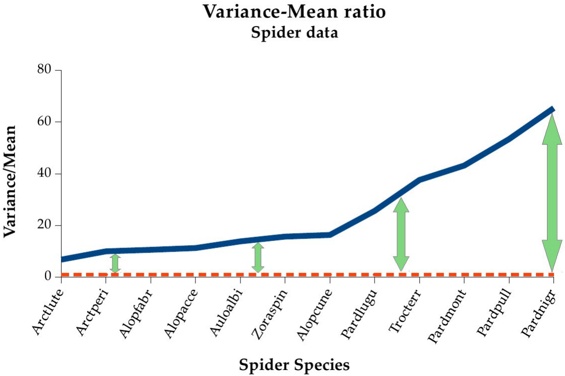
$$y_{ij} = \begin{cases} (0)\ \textbf{None} & \text{No data recorded} \\ (1)\ \textbf{Low} & \text{Species coverage is below 25\%} \\ (2)\ \textbf{Medium} & \text{Species coverage is between 25\% -- 65\%} \\ (3)\ \textbf{High} & \text{Species coverage is higher than 65\%} \end{cases}$$

Table : Frequencies of spider abundance by site, in 4-level ordinal scale.

| **Ordinal scale** | **0** | **1** | **2** | **3** | **Total** |
|---|---|---|---|---|---|
| Spider abundance | No data recorded | Low | Medium | High | Total |
| Frequency ($y_{ij}$) | 154 | 66 | 56 | 60 | 336 |

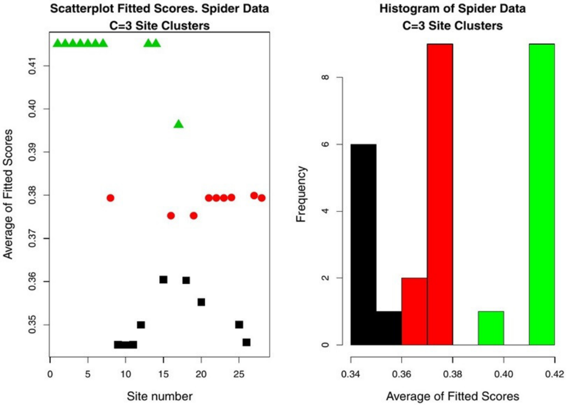
# 4. Application. Spider Dataset

Blue line: Variance-mean ratio (sorted ascending) for the spider data set.
Orange dashed line: indicates no overdispersion.
Green arrows: **Overdispersion** (variance$>$mean) is observed in all the species.



# 4. Application. Spider Dataset

Scatter plot and histogram of the $R = 3$ fitted sites clusters $\{\overline{\phi}_{(i.)}\}$ from the row clustering version of the stereotype model $(\mu_k + \phi_k(\alpha_r + \beta_j))$.



# Cluster Results
# Count data vs. Ordinal data

# 3. Result Comparison: Count data vs. Ordinal data

Table : **Spider data set**: Site clustering results for Poisson, NB and ordered stereotype model.

| Groups | **Clustering (highest probability)** | | |
|---|---|---|---|
| | **Poisson** | **NB** | **Stereotype** |
| R1 | {**1-7**,9-12, **13, 14**, 25} | {**1-7,13,14**} | {**1-7,13,14**} |
| R2 | {22-24,26-28} | {9-12,22-28} | {8,21-24,27,28} |
| R3 | {8,15-21} | {8,15-21} | {9-12,15-20,25-26} |

# 3. Results Column Clustering. Comparison

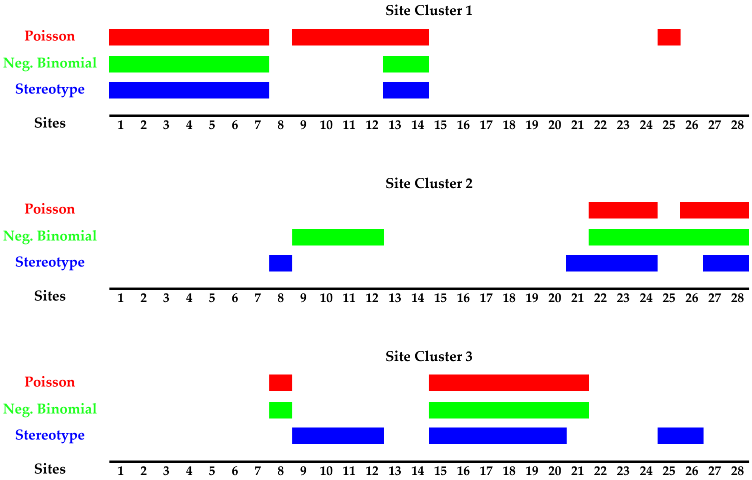



Figure : **Spider data C=3**: Poisson, Neg. Bin. and Ordinal Stereotype

# 3. Results Column Clustering. Comparison

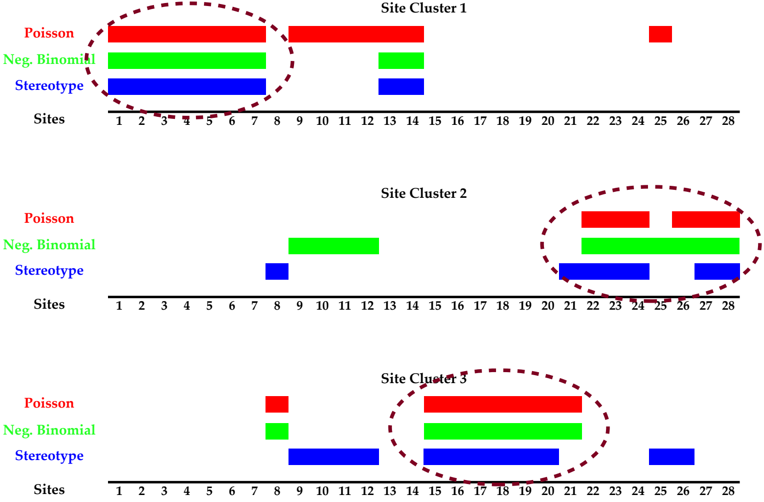



Figure : **Spider data C=3**: Poisson, Neg. Bin. and Ordinal Stereotype

# 3. Results Column Clustering. Comparison

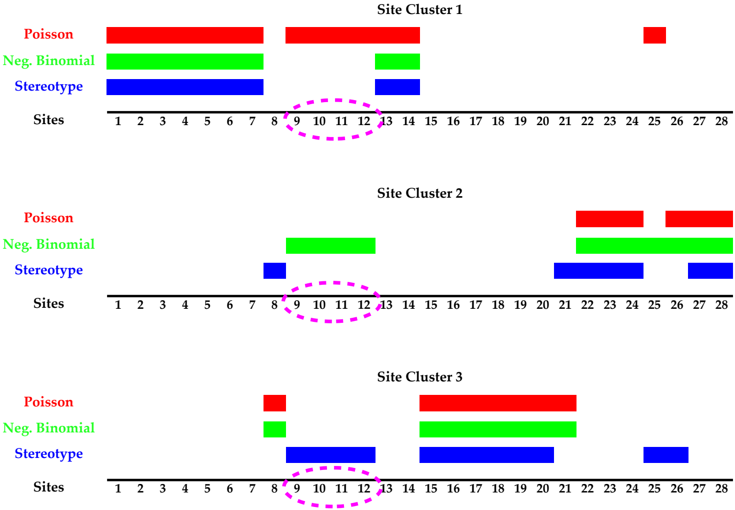



Figure : **Spider data C=3**: Poisson, Neg. Bin. and Ordinal Stereotype

# 3. Results Column Clustering. Comparison

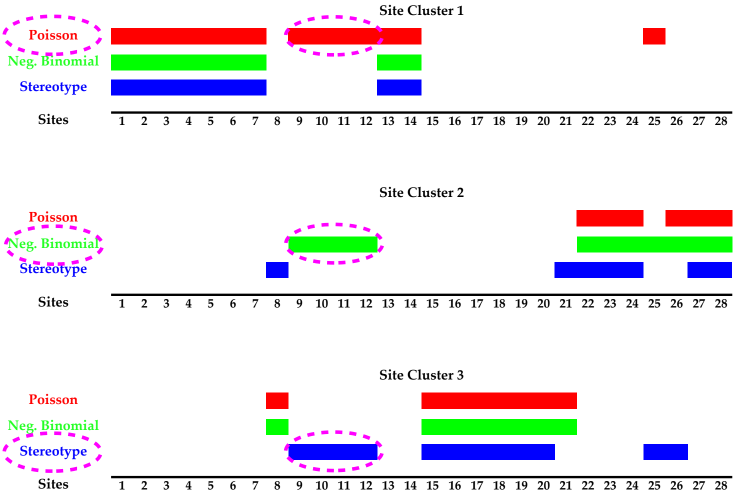



Figure : **Spider data C=3**: Poisson, Neg. Bin. and Ordinal Stereotype

# 3. Results Column Clustering. Comparison

- Clustering measures:
  - Variation of information (VI, Meila (2005)).
  - Normalized information distance (NID, Kraskov *et al.* (2005)),
  - Adjusted Rand index (ARI, Hubert *et al.* (1985))
- Range between (0,1).
- Large values indicate similarity of clustering.

Table : **Spider data set**: Clustering results for Poisson, NB, and stereotype model. Stereotype is closer to NB than Poisson.

| Clustering Comparison | ARI | 1-NVI | 1-NID |
|---|---|---|---|
| Poisson vs. Stereotype | 0.334 | 0.304 | 0.457 |
| **NB vs. Stereotype** | **0.465** | **0.423** | **0.590** |

# 5. Summary. Conclusions

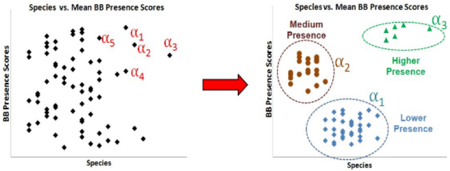
- Features of categorising count data into ordinal data were shown.
- In our view, advantages:
  - We **do not have** to decide among different parametric models for the data. (i.e. it enables the inclusion of all of the different levels of dispersion in one methodology.)
  - Replacing high/low counts with "high/low" ordinal categories makes the actual **counts less influential** in the model fitting.
  - **Saving in cost** of sampling time in collecting only ordinal data (sample more sites).
- Future research directions:
  - Numerical experiment: Investigate the differences between recoded and original count data.
  - Developing a measure to quantify the loss of information.

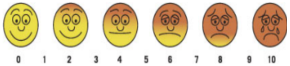
# Acknowledgments and References

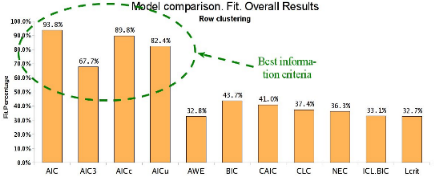
- Shirley Pledger and Richard Arnold.
- Funding: Victoria University of Wellington.

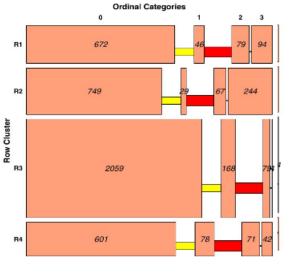
1. Rogers, A. Statistical Analysis of Spatial Dispersion: The Quadrat Method. Monographs in Spatial and Environmental Systems Analysis. Pion, 1974.
2. Warton, D. I., Wright, S. T., and Wang, Y. Distance-based multivariate analyses confound location and dispersion effects. Methods in Ecology and Evolution, 3(1):89-101, 2012.
3. Pledger, S. and Arnold, R. Multivariate methods using mixtures: Correspondence analysis, scaling and pattern-detection. Computational Statistics and Data Analysis, 71:241-261, 2014.
4. Fernández, D., Arnold, R., and Pledger, S. Mixture-based clustering for the ordered stereotype model. Computational Statistics and Data Analysis, 2015.
5. Fernández, D. and Pledger, S. Categorising Count Data into Ordinal Responses with Application to Ecological Communities. JABES, (forthcoming 2016).

# Thanks for listening!!!



# References

## Questions?


1. Rogers, A. Statistical Analysis of Spatial Dispersion: The Quadrat Method. Monographs in Spatial and Environmental Systems Analysis. Pion, 1974.

2. Warton, D. I., Wright, S. T., and Wang, Y. Distance-based multivariate analyses confound location and dispersion effects. Methods in Ecology and Evolution, 3(1):89-101, 2012.

3. Pledger, S. and Arnold, R. Multivariate methods using mixtures: Correspondence analysis, scaling and pattern-detection. Computational Statistics and Data Analysis, 71:241-261, 2014.

4. Fernández, D., Arnold, R., and Pledger, S. Mixture-based clustering for the ordered stereotype model. Computational Statistics and Data Analysis, 2015.

5. Fernández, D. and Pledger, S. Categorising Count Data into Ordinal Responses with Application to Ecological Communities. JABES, (forthcoming 2016).

# Extra Slides

# 1. Approaches. Ordinal stereotype model

For example, **Row clustered ordinal stereotype model**:

$$\log\left(\frac{P\left[y_{ij}=k \mid i \in r\right]}{P\left[y_{ij}=1 \mid i \in r\right]}\right)=\mu_k+\phi_k(\alpha_r+\beta_j)$$

$$i=1,\dots,n \quad j=1,\dots,p \quad k=1,\dots,q \quad r=1,\dots,R<n$$

- $\mu_k$: cut points (nuisance parameters).
- $\alpha_r$: effect of the row cluster $r$.
- $\beta_j$: effect of the columns.
- $\phi_k$: "score" for the response category k.

- Including an increasing order constraint:

$$0=\phi_1 \leq \phi_2 \leq \dots \leq \phi_q = 1 \ ,$$

captures the ordinal nature of the outcomes.