The statistical analysis of compositional data:
Data, scale and random variables

Prof. Dr. Juan José Egozcue
Prof. Dr. Vera Pawlowsky-Glahn
Ass. Prof. Dr. René Meziat

Instituto Colombiano del Petróleo
Piedecuesta, Santander, Colombia
March 20–23, 2007

Summary

Kinds of data

- Experiments produce results which are observed.
- Observations are measured and/or classified
    - If classified: qualitative or categorical data. No order is implied.
    - If measured: quantitative data. Order is implicit.
        - Discrete measurements (all measurements)
        - Continuous measurements (an assumption)

Measures of difference (positive case)

Example: tallness of a person

Two adults are $a_1 = 160$ and $a_2 = 180$ cm tall.
Two babies are $b_1 = 40$ and $b_2 = 60$ cm tall

- Are the two differences 20cm? (Absolute scale)
- Or better: the first adult is $a_1/a_2 = 0.89$ times the second, and the first baby is $b_1/b_2 = 0.67$ (Relative scale)

For the relative scale symmetry would be preferable:

$$\frac{a_1}{a_2} - \frac{b_1}{b_2} = 0.22 \neq 0.38 = \frac{b_2}{b_1} - \frac{a_2}{a_1}$$

- Log-ratio gives symmetry to relative scale:

$$\ln(a_1) - \ln(a_2) = -0.12 \;,\; \ln(b_1) - \ln(b_2) = -0.41$$

Measures of difference (interval case)

Probabilities of an event: How do you measure differences between probabilities?

- Absolute: $|p_2 - p_1|$
- Relative: $|\ln(p_2) - \ln(p_1)|$
- Logistic: $|\ln(p_2/(1-p_2)) - \ln(p_1/(1-p_1))|/\sqrt{2}$

p_1	p_2	abs. dif.	rel. dif.	$\mathcal{S}^2$ dif.
0.0001	0.0002	0.0001	0.6931	0.4902
0.5	0.5001	0.0001	0.0002	0.0003
0.9999	0.9998	0.0001	0.0001	0.4902

- Absolute: no scale at all
- Relative: scaled near 0; no symmetry
- Logistic: scaled; symmetry

Behavior at the borders or end-points

Examples:

- Tallness of 0cm does not correspond to a person!
- If an event has prob. 0 or 1, the probabilistic study is useless!
- If there is exactly 0ppb of an element, please forget it!
- An earthquake of magnitude 0 is not an earthquake!
- A temperature of 0 Kelvin is unattainable!

Absurd or unattainable border points:

They should be placed at the infinity of the scale!

Behavior at the borders or end-points

Examples:

- Tallness of 0cm does not correspond to a person!
- If an event has prob. 0 or 1, the probabilistic study is useless!
- If there is exactly 0ppb of an element, please forget it!
- An earthquake of magnitude 0 is not an earthquake!
- A temperature of 0 Kelvin is unattainable!

Absurd or unattainable border points:

They should be placed at the infinity of the scale!

Scaling transformations of data

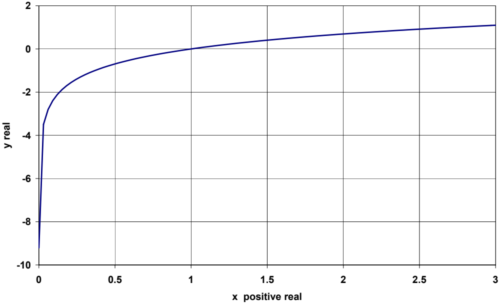
Cases of data support and recommended transformation:

- Real data: scale is absolute. No transformation.
  Examples: unknown!
- Positive data: scale is relative. Log-transformation
  Examples: Wind speed, wave-height, earthquake magnitude, tallness...
- Interval data: scale is relative and symmetric. Logistic transformation.
  Examples: proportions, concentrations, probabilities...

Logarithmic transformation

Log-transformation: $\mathbb{R}_+ \to \mathbb{R}$

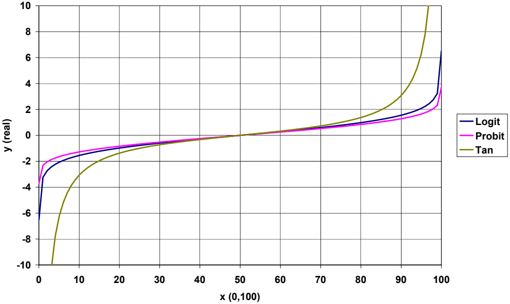
$$y = \ln(x - x_0) \quad , \quad x_0 = 0$$



Logistic (Logit) transformation

Logit transformation $(a, b) \to \mathbb{R}$

$$y = \ln \frac{x-a}{b-x} \quad , \quad a = 0,\ b = 100$$



Transformations of data

- They do change the scale of data
- The adequate transformation is a subjective choice. It depends on
  - information carried by the data (relative, absolute, directional,...)
  - how differences are measured (ratios, differences,...)
  - the support of the observations (real, positive, interval,...)
- Adjustment to a given distribution is not a good reason for transformation of data

Transformations of data

- They do change the scale of data
- The adequate transformation is a subjective choice. It depends on
  - information carried by the data (relative, absolute, directional,...)
  - how differences are measured (ratios, differences,...)
  - the support of the observations (real, positive, interval,...)
- Adjustment to a given distribution is not a good reason for transformation of data

Random variables

$$X : \underbrace{(\Omega, \mathfrak{A}, P_\Omega)}_{\substack{\color{red}{\text{probabilityspace}} \\ \color{blue}{\text{experimental conditions}}}} \underbrace{\rightarrow}_{\substack{\color{red}{\text{random variables}} \\ \color{blue}{\text{experiment}}}} \underbrace{(S, \mathfrak{B}, P)}_{\substack{\color{red}{\text{sample space}} \\ \color{blue}{\text{observable results}}}}$$

Events: $\mathfrak{A}$, $\mathfrak{B}$

$$B \in \mathfrak{B} \;\Rightarrow\; X^{-1}(B) = A \in \mathfrak{A}$$

Probabilities: $X = X(\omega)$

$$\mathrm{P}[X \in B] = \mathrm{P}_\Omega[\omega \in A]$$

Univariate random variable X, sample space $\mathbb{R}$

Cumulative distribution function (cdf)

$$F_X(x) = \mathrm{P}[X \leq x], X \in \mathbb{R}$$

Probability function: support is discrete

$$p_X(x_i) = \mathrm{P}[X = x_i] = F_X(x_{i+1}) - F_X(x_i)\ ,\ x_i \in \text{support}$$

Probability density function(pdf) support $S \in \mathbb{R}$

$$B \in \mathfrak{B}\ ,\ B \subset \mathbb{R}\ ,\ \mathrm{P}[X \in B] = \int_B f_X(x)\, dx$$

$$f(x) = \frac{d}{dx} F_X(x) \quad \text{(a.e.)}$$

Disappointing issues about pdf (I)

Assume that the support is $X > 0$. Then

$$\mathrm{P}[X = x_0 > 0] = \int_{\{x_0\}} f_X(x)\, dx = 0 \quad !!!$$

Which is the difference between $x_0 > 0$, a possible value of X, and an impossible value $x_1 < 0$, also satisfying $\mathrm{P}[X = x_1 < 0] = 0$?

The sample space $\mathbb{R}$ is not adequate

Try to stretch the sample space to $\mathbb{R}_+$!

Disappointing issues about pdf (I)

Assume that the support is $X > 0$. Then

$$\mathrm{P}[X = x_0 > 0] = \int_{\{x_0\}} f_X(x)\, dx = 0 \quad !!!$$

Which is the difference between $x_0 > 0$, a possible value of X, and an impossible value $x_1 < 0$, also satisfying $\mathrm{P}[X = x_1 < 0] = 0$?

The sample space $\mathbb{R}$ is not adequate

Try to stretch the sample space to $\mathbb{R}_+$!

Disappointing issues about pdf (II)

Probability is represented by definition

$$\mathrm{P}[X \in B] = \int_B d\mathrm{P} = \int_B \frac{d\mathrm{P}}{d\lambda} d\lambda \approx \sum \frac{d\mathrm{P}}{d\lambda}(b_i') \cdot \lambda\{b_i, b_{i+1}\}$$

pdf with respect to the reference measure λ

$$\frac{d\mathrm{P}}{d\lambda}(x) = f_X^{\lambda}(x)$$

Why $\lambda\{b_i, b_{i+1}\} = |b_{i+1} - b_i|$ (Lebesgue measure)?

Selection of the reference measure λ:

should be in accordance of the scale of data!

And the pdf depends on λ!!!

Disappointing issues about pdf (II)

Probability is represented by definition

$$\mathrm{P}[X \in B] = \int_B d\mathrm{P} = \int_B \frac{d\mathrm{P}}{d\lambda} d\lambda \approx \sum \frac{d\mathrm{P}}{d\lambda}(b_i') \cdot \lambda\{b_i, b_{i+1}\}$$

pdf with respect to the reference measure λ

$$\frac{d\mathrm{P}}{d\lambda}(x) = f_X^{\lambda}(x)$$

Why $\lambda\{b_i, b_{i+1}\} = |b_{i+1} - b_i|$ (Lebesgue measure)?

Selection of the reference measure λ:

should be in accordance of the scale of data!

And the pdf depends on λ!!!

Mean and variance

X r.v. with pdf f_X (sample space $\mathbb{R}$)

Mean

$$\mathrm{E}[X] = \mu = \int_{\mathbb{R}} x\, f_X(x)\, dx$$

Variance

$$\mathrm{Var}[X] = \int_{\mathbb{R}} (x-\mu)^2\, f_X(x)\, dx \;,\;\; \mu = \mathrm{E}[X]$$

Alternative definitions

Variability: $V(\xi) = \int_{\mathbb{R}} d^2(x,\xi)\, f_X(x)\, dx$

Mean: $\mu = \mathrm{argmin}_\xi V(\xi)$

Variance: $\mathrm{Var}[X] = V(\mu) = \int_{\mathbb{R}} d^2(x,\mu)\, f_X(x)\, dx$

Mean and variance

X r.v. with pdf f_X (sample space $\mathbb{R}$)

Mean

$$\mathrm{E}[X] = \mu = \int_{\mathbb{R}} x \, f_X(x) \, dx$$

Variance

$$\mathrm{Var}[X] = \int_{\mathbb{R}} (x-\mu)^2 \, f_X(x) \, dx \ , \ \mu = \mathrm{E}[X]$$

Alternative definitions

Variability: $V(\xi) = \int_{\mathbb{R}} d^2(x,\xi) \, f_X(x) \, dx$

Mean: $\mu = \mathrm{argmin}_{\xi} V(\xi)$

Variance: $\mathrm{Var}[X] = V(\mu) = \int_{\mathbb{R}} d^2(x,\mu) \, f_X(x) \, dx$

Disappointing issues about mean and variance

Distance

In $\mathbb{R}$, $d(x, y) = |x - y|$

In $\mathbb{R}_+$, $d_+(x, y) = |\ln x - \ln y|$

If the sample space is $\mathbb{R}_+$,

Why to use $d(\cdot, \cdot)$ of $\mathbb{R}$? Do these definitions work better?

$$\mathrm{Var}_+[X] = \int_{\mathbb{R}_+} d_+^2(x, \mu)\, f_X(x)\, dx$$

$$\mathrm{E}_+[X] = \int_{\mathbb{R}_+} x\, f_X^{\lambda_+}(x)\, d\lambda_+ = \exp(\mathrm{E}[\ln X])$$

$$\lambda_+\{a, b\} = |\ln b - \ln a|$$

Normal distribution (univariate)

Sample space and support: $\mathbb{R}$

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- $\mathrm{E}[X] = \mu \quad , \quad \mathrm{Var}[X] = \sigma^2$
- Symmetric with respect to μ
- Sums of normal variables are normal
- Sums of non-normal variables are approached by normal ones

Multivariate normal distribution

Sample space and support: $\mathbb{R}$

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{2\pi(\det\Sigma)^n}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

- $\mathrm{E}[\mathbf{X}] = \boldsymbol{\mu}$
- $\mathrm{Cov}[\mathbf{X}] = \mathrm{E}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})'] = \Sigma$, $(\det\Sigma \neq 0)$
- Symmetric with respect to $\boldsymbol{\mu}$
- Sums, marginals and conditionals of normal variables are normal
- Sums of non-normal variables are approached by normal ones