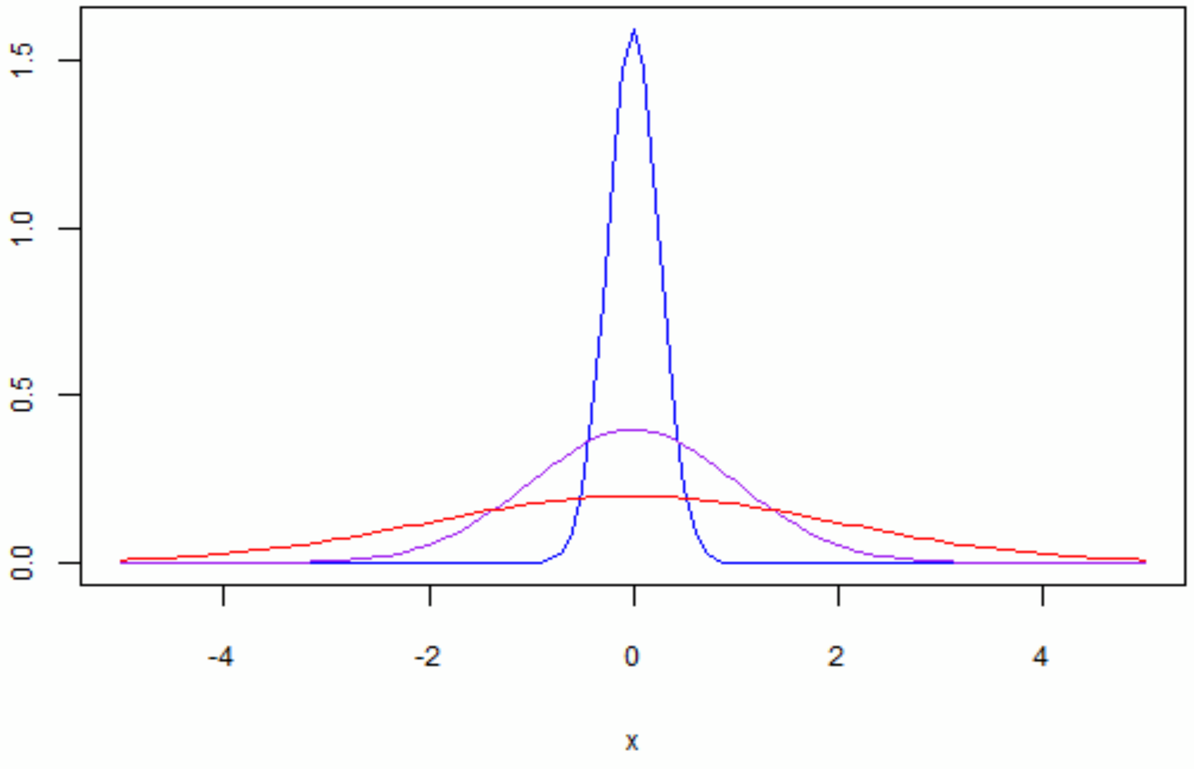
Dispersion or Variability

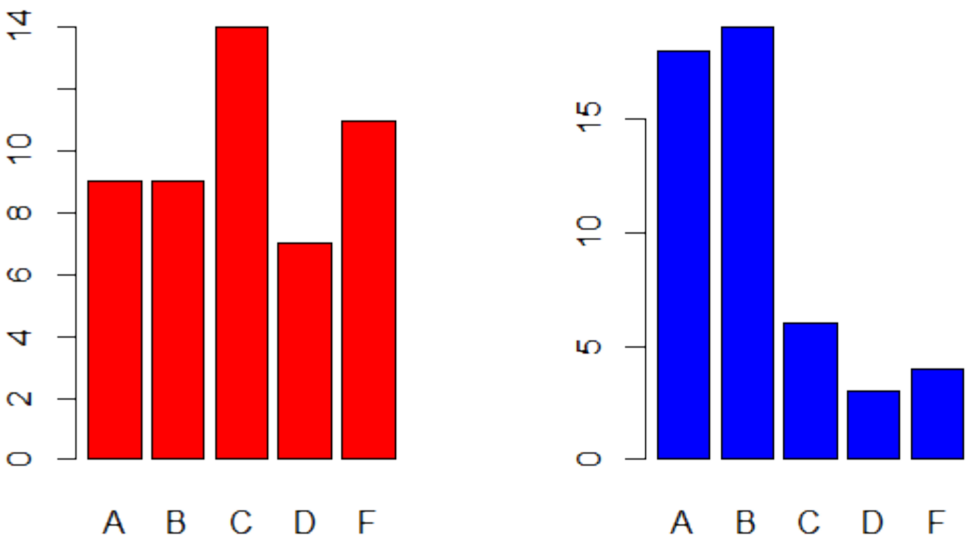
How Much Do Distributions or Data Vary?

Charles Peters
University of Houston

More Variable – Less Variable
Numeric



More Variable – Less Variable
Categorical



Variability of Numeric Variables

- Variance: $\sigma^2 = E[(X-\mu)^2]$, $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- Standard deviation: $\sigma = \sqrt{\sigma^2}$, $s = \sqrt{s^2}$
- Interquartile range: $IQR = quantile(X, .75) - quantile(X, .25)$
- Median absolute deviation: $MAD = median\{|X - m|\}$, $m = median(X)$

Robustness

- IQR and MAD are *robust* measures of variability. Insensitive to a few outliers.
- Standard deviation is not robust. One extreme outlier can change its value drastically.
- All are *scale* parameters or statistics. When the scale of measurement is changed, they change in the same way.

Variability of Categorical Variables Multinomial Distributions

- A categorical variable has m possible values, with probabilities $p_1, \cdots, p_m$, positive and summing to 1.
- Replicate the experiment N times independently. Possibly $N=1$.
- $Y_i =$ number of occurrences of i^{th} outcome.
- This is a *multinomial experiment* and the random vector $Y=(Y_1, \cdots, Y_m)$ has a multinomial distribution.

Gini Measure of Variability

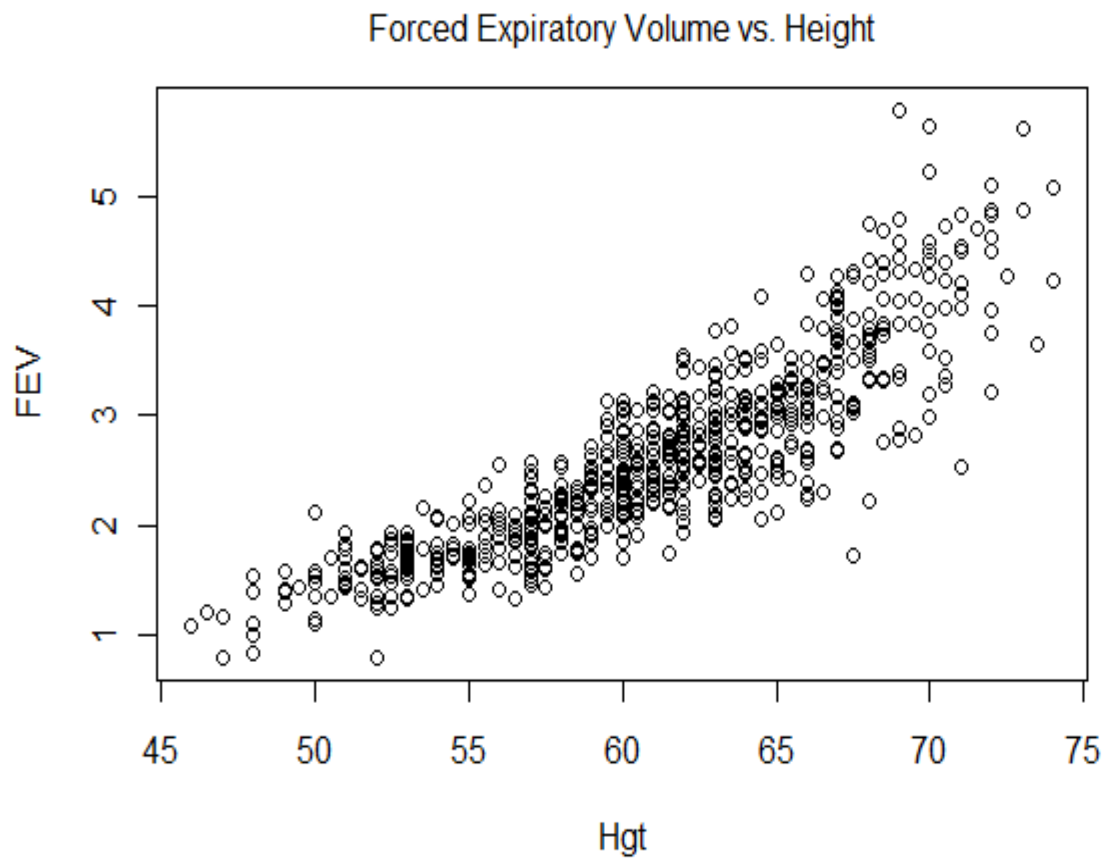
- In the multinomial distribution, each component Y_i has a binomial distribution with variance $Np_i(1-p_i)$.
- $Gini = N\sum_{1}^{m} p_i(1-p_i) = N(1-\sum_{1}^{m} p_i^2)$
- Since $\sum_{1}^{m} p_i = 1$, $Gini$ is maximum when each $p_i = 1/m$, i.e., all category levels are equally likely, and 0 when some $p_i = 1$, others = 0.
- Note: The maximum value increases with m.
- With data, replace p_i by its estimate Y_i/N.

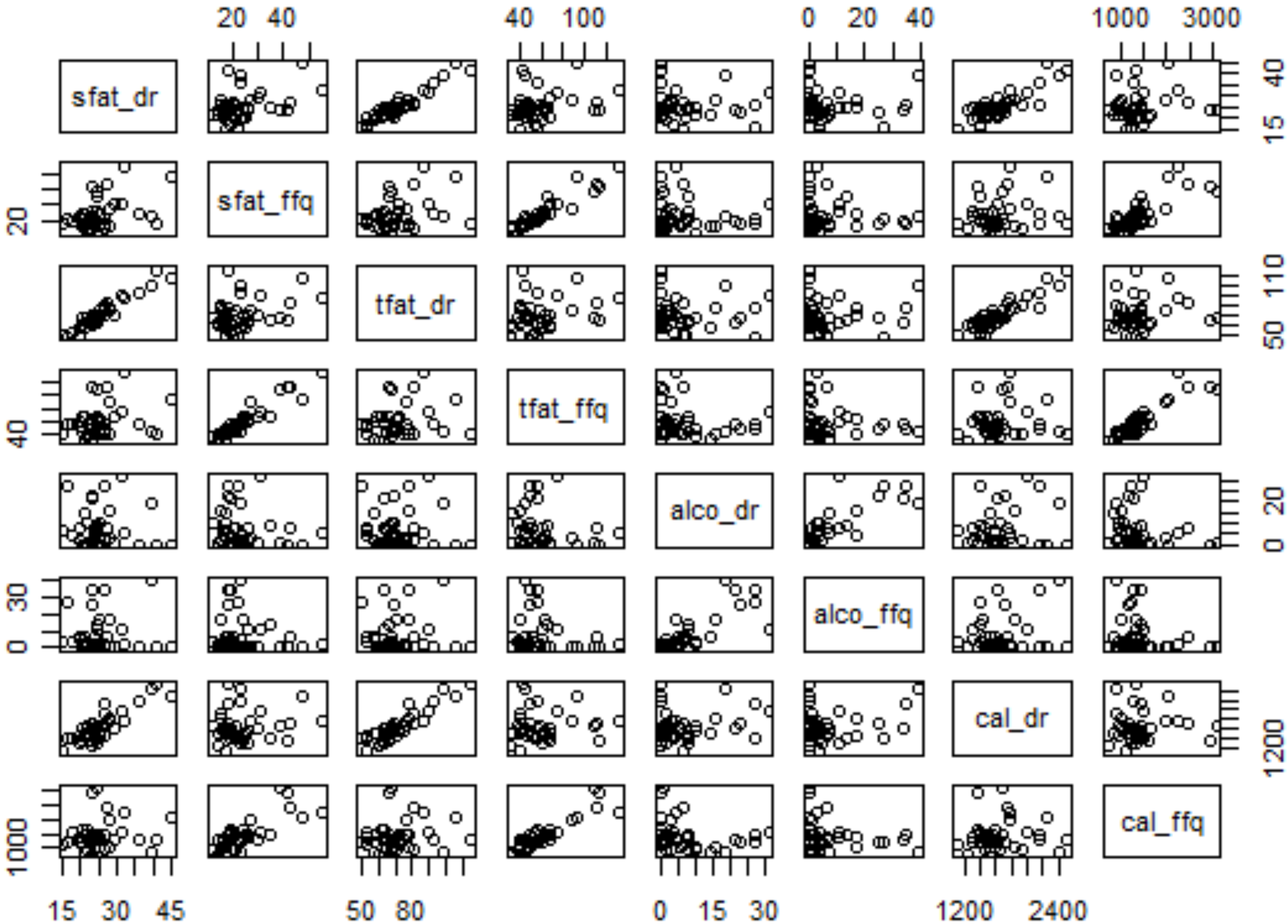
Entropy Measure of Variability

- $H = -N \sum_{1}^{m} p_i \log p_i$
- By continuity, define $0 \log 0 = 0$. Then $0 \le H \le N \log m$.
- $H = 0$ when some $p_i = 1$. $H = N \log m$ when all $p_i = 1/m$. The maximum value increases with m.
- With data, replace p_i by its estimate Y_i / N. Then H is related to the likelihood ratio statistic for the null hypothesis of equally likely category levels.

Correlation

To What Extent Are Variables Related?


Forced Expiratory Volume vs. Height
FEV
1
2
3
4
5
45
50
55
60
65
70
75
Hgt


sfat_dr
sfat_ffq
tfat_dr
tfat_ffq
alco_dr
alco_ffq
cal_dr
cal_ffq
20 40
40 100
0 20 40
1000 3000
15 30 45
50 80
0 15 30
1200 2400
40
15
20
110
50
40
20
0
30
0
1200
1000

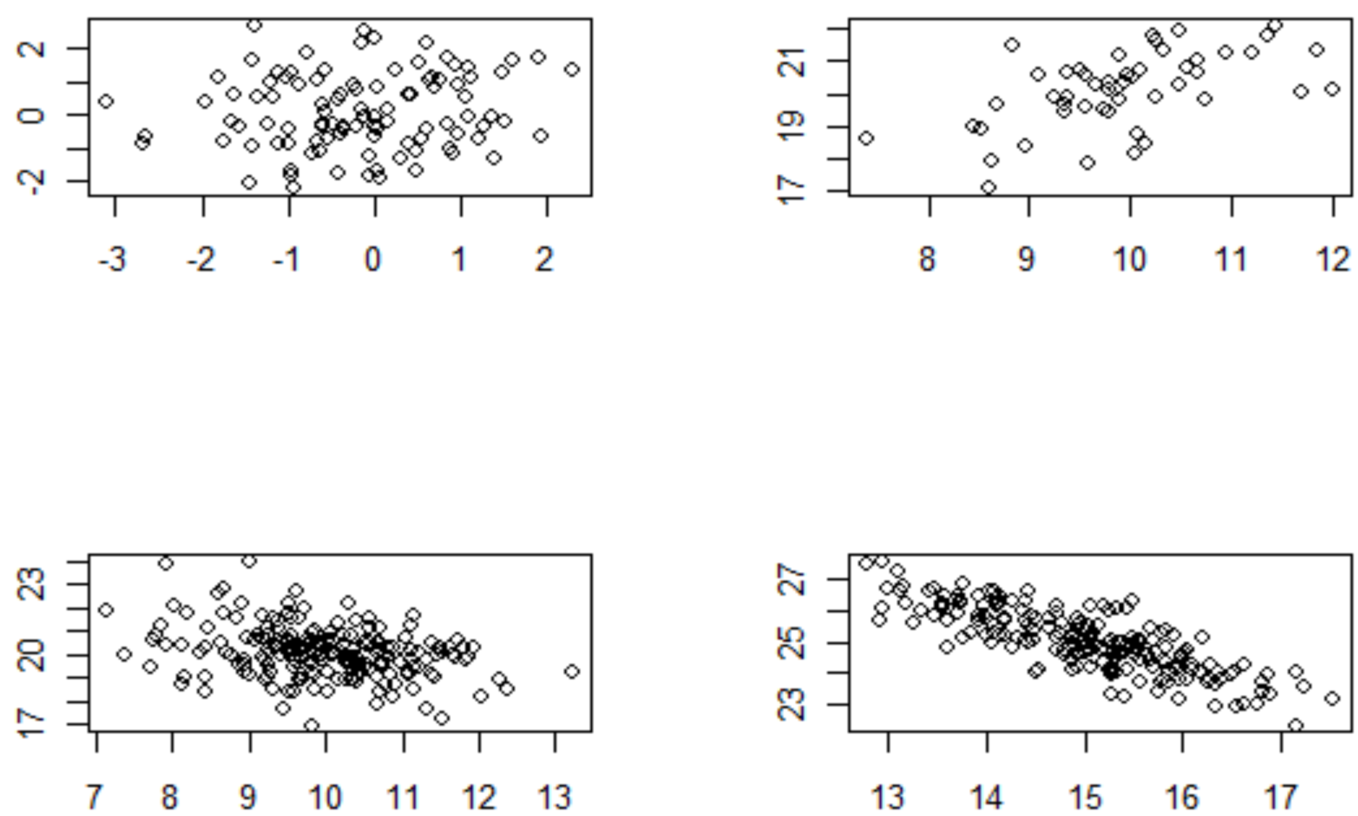
Theoretical Covariance and Correlation Pearson Correlation

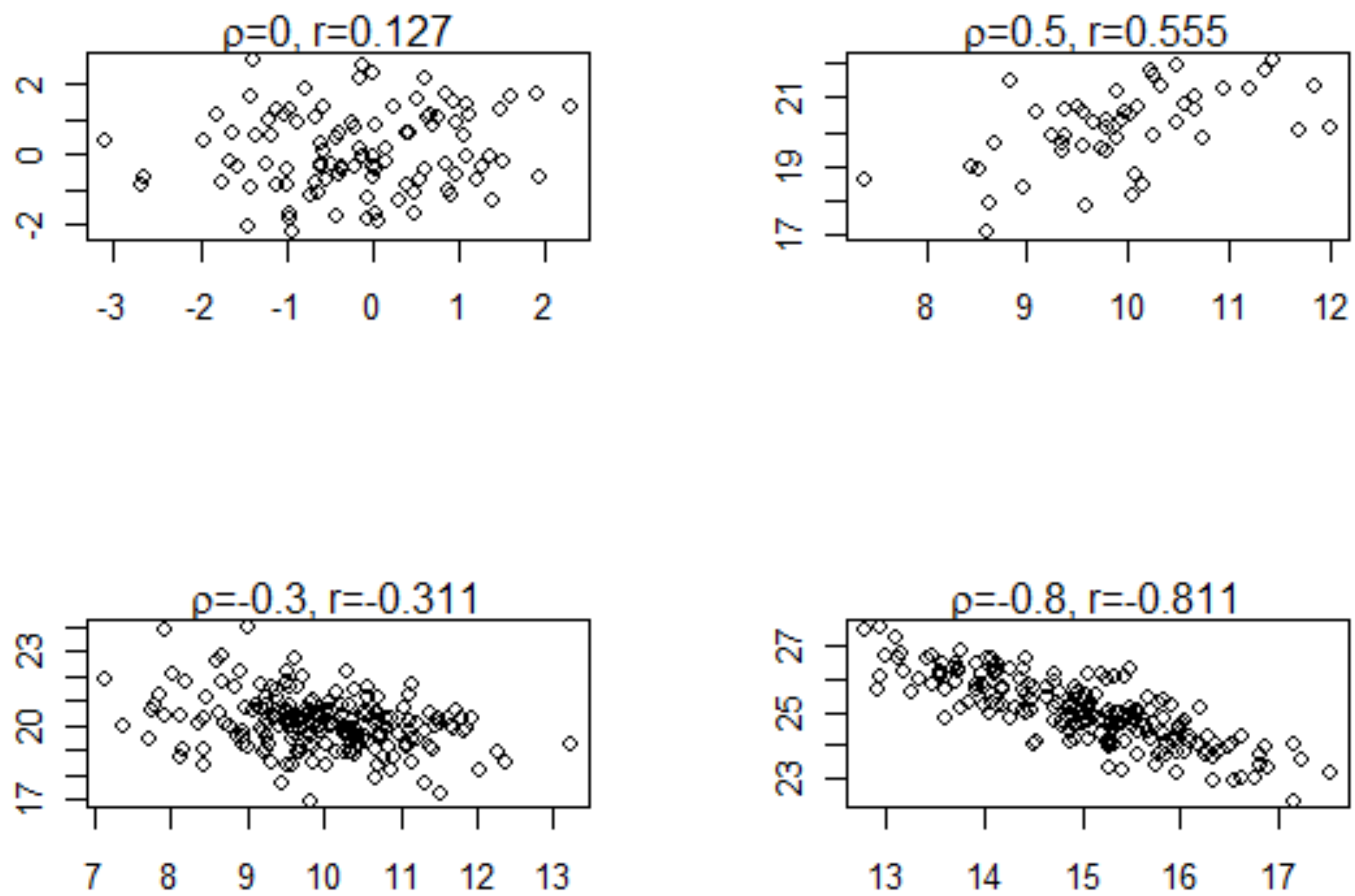
- X,Y jointly distributed random variables
- Means μ_x, μ_y, standard deviations $\sigma_x > 0$, $\sigma_y > 0$.
- $cov(X,Y) = E[(X - \mu_x)(Y - \mu_y)]$
- $cor(X,Y) = \rho_{xy} = cov(X,Y)/\sigma_x \sigma_y$
- $|\rho| \leq 1$, with equality iff $aX + bY = c$ for constants a, b, c.

Sample Covariance and Correlation

- $s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$
- $s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - x)^2$
- $s_y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - y)^2$
- $r_{xy} = s_{xy} / s_x s_y$
- Random variables. $|r_{xy}| \leq 1$ with equality iff $a x_i + b y_i = c$ for all i.

Guess ρ, Guess r





ρ=0, r=0.127
ρ=0.5, r=0.555
ρ=-0.3, r=-0.311
ρ=-0.8, r=-0.811

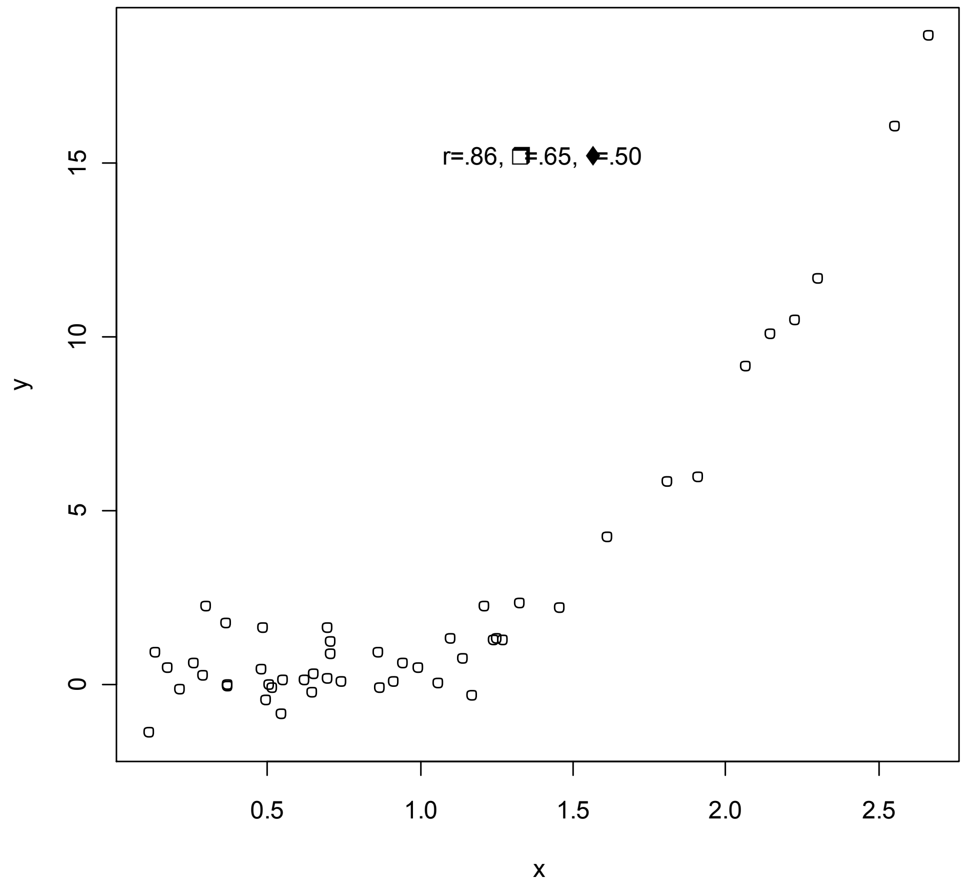
Spearman's Rho

- Given data $(x_1, y_1), \cdots (x_n, y_n)$, rank the x's and also rank the y's. Let $u_i = rank(x_i)$ and $v_i = rank(y_i)$.
- Then calculate the Pearson correlation of the pairs $(u_1, v_1), \cdots, (u_n, v_n)$.
- This is Spearman's rho ρ_s.
- If X and Y are independent, the distribution of ρ_s does not depend on their distributions.
- Provides a nonparametric or distribution-free test of no association between X and Y.

Kendall's Tau

- Count the number c of pairs of indices (i,j) with $i<j$ and $(x_i - x_j)(y_i - y_j) > 0$. These are *concordant pairs*.
- The number d of discordant pairs is $p - c$, where $p = 1/2\, n(n-1)$.
- $\tau = c - d / p$
- τ is distribution free if X and Y are independent.

Comparison



Variance-Covariance Matrices

Random Vectors

Variance-Covariance Matrix

- $X_1, X_2, \cdots, X_m$ jointly distributed numeric variables.
- $X = (X_1, X_2, \cdots, X_m)^t \in R^{m \times 1}$ is a random vector.
- $V = V(X) = (v_{ij}) \in R^{m \times m}$, where $v_{ij} = cov(X_i, X_j) = \rho_{ij}\sigma_i\sigma_j$, $\rho_{ij} = cor(X_i, X_j)$.
- Positive definite, symmetric matrix with positive eigenvalues, orthogonal eigenvectors.
- Given n sample observations of X, the sample variance-covariance matrix V has sample correlations and standard deviations.

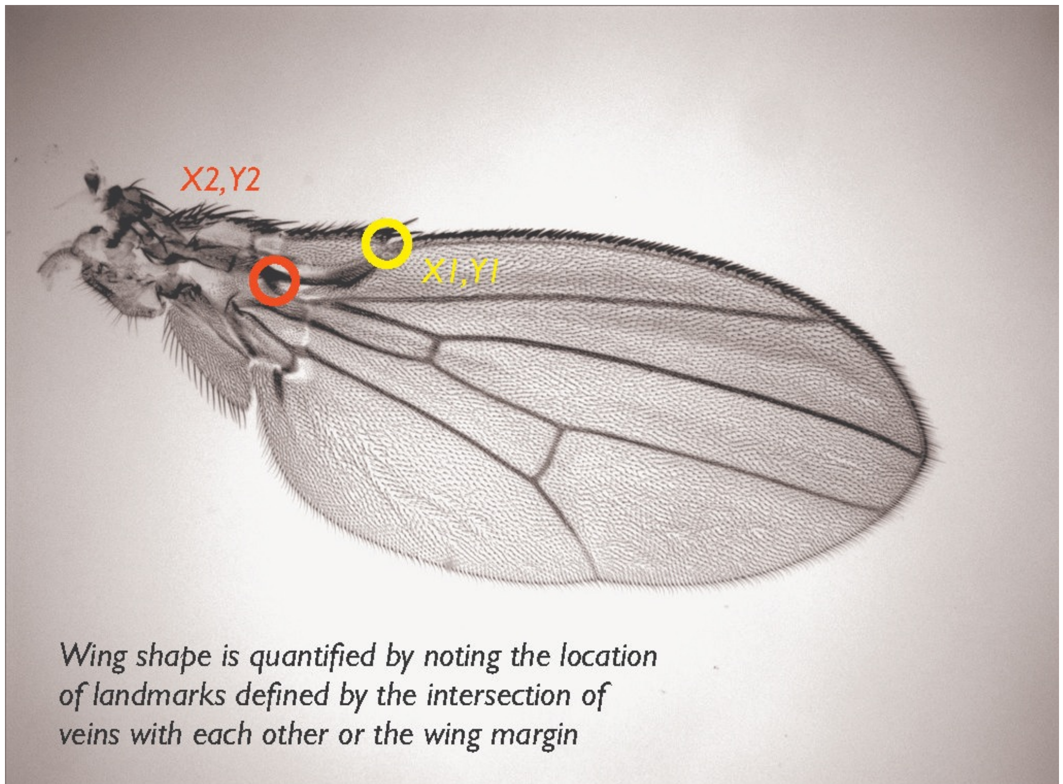
Principal Components

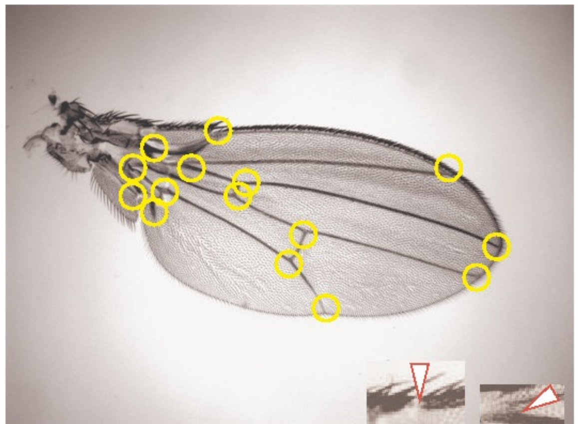
- $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m > 0$ the ordered eigenvalues of V.
- $u_1, u_2, \cdots, u_m$ corresponding orthogonal unit eigenvectors.
- $u_1 \cdot X$, $u_2 \cdot X, \cdots, u_m \cdot X$ are <u>uncorrelated</u>. Called the principal components of the random vector X.
- $\lambda_1 = var(u_1 \cdot X)$, $\lambda_2 = var(u_2 \cdot X)$, etc.

Importance of Principal Components

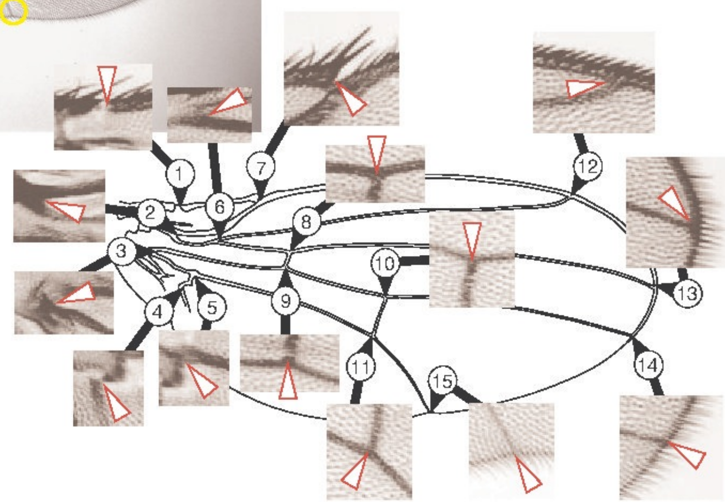
- $\sum_{i=1}^{m} \lambda_i = \sum_{i=1}^{m} var(X_i)$
- If the first few largest λ_i strongly dominate, most of the variation of the random vector X is captured by the first few principal components.
- Useful as a dimensionality reduction tool.

Fruit Fly Wing Shape
Courtesy Prof. Tony Frankino BIOL/ BCHS

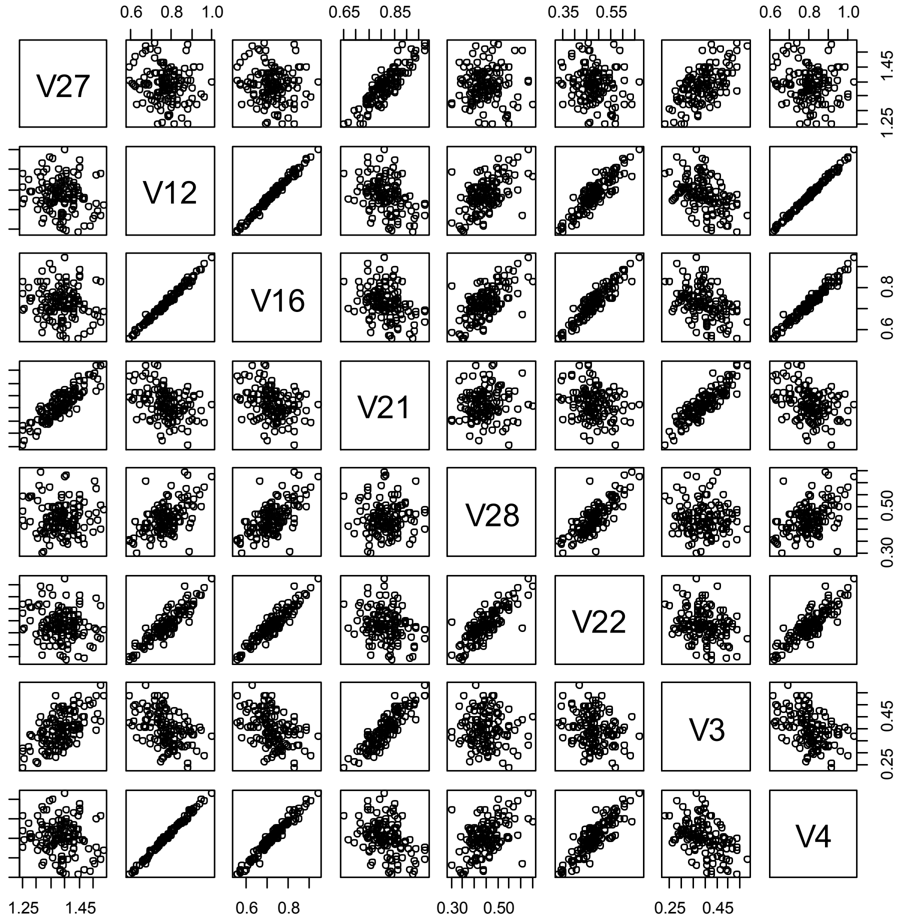




Total of 15 landmarks used.



Some of the Variables

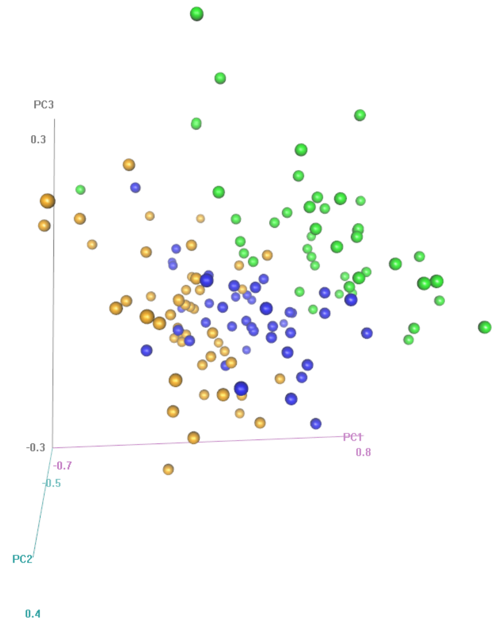


Variances of Principal Components

[1]   0.0922 0.0405 0.0116 0.0054 0.0004 0.0002 0.0001 0.0001 0.0000 0.0000
[11] 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
[21] 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000

Total variation is 0.15. Top three carry most of it.

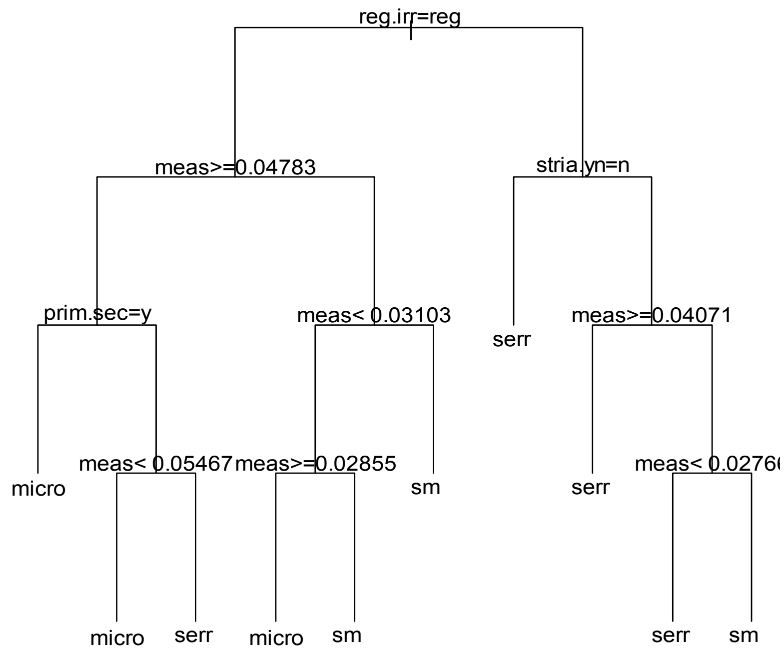
Principal Components by Species



Classification Trees

- Splitting of nodes always decreases Gini or entropy. So splits always increase "purity" of terminal nodes.
- Split nodes on single variables, nodes and variables chosen to maximize the decrease in total Gini or entropy.
- Stop when the decrease falls below a threshold or when nodes get too small.

Example



reg.irr=reg
meas>=0.04783
stria.yn=n
prim.sec=y
meas< 0.03103
serr
meas>=0.04071
micro
meas< 0.05467
meas>=0.02855
sm
serr
meas< 0.0276
micro
serr
micro
sm
serr
sm