Overview

Real statistical application:
- Remote monitoring of east African long rains
- Lead up to Lab 5-6

Review of bivariate/multivariate relationships
- Definition of variance
- Definition of co-variance

Extension to multi-variate case
East African Food Insecurity

FEWS NET
Food Insecurity Severity Scale
- No Acute Food Insecurity
- Moderately Food Insecure
- Highly Food Insecure
- Extremely Food Insecure
- Famine

C. Funk Geog 210C Spring 2011
Maize: Nominal wholesale prices in Nairobi

| JAN | FEB | MAR | APR | MAY | JUN | JUL | AUG | SEP | OCT | NOV | DEC |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
|     |     |     |     |     |     |     |     |     |     |     |     |     |

- 5 year average 2006-2010
- Previous year 2010
- Current year 2011
Child Stunting

Map showing the percentage of children under 5 years old who are stunted in various regions of East Africa, with particular emphasis on Kenya. The map highlights areas with high percentages of stunted children, indicating regions with significant nutritional challenges.
Kenya Population and Number of raindays, past 30 days as of April 17th

Image on the left shows Landscan population density overlain with a blue mask. Regions not masked are in Kenya and had less than 9 rain days during the past month.

Halfway through the season, a large population center appears to be at risk?

Percent of March-April Rainfall

Average March-April rainfall as a percent of March-June totals

- >90%
- >80%
- >70%
- >60%
- >60%
- <60%
Concerns for Central & Eastern Kenya

If the rest of the seasonal is normal (which is probably unlikely), central Kenya will have seasonal totals 24% below normal. If the season is plays out like 2009, Central Kenya might receive ~40% below normal.

If the rest of the seasonal is normal (which is probably unlikely), central Kenya will have seasonal totals 23% below normal. If the season is plays out like 2009, Central Kenya might receive ~50% of normal.

Area NE of Nairobi has had +6-7°C LST Anomalies, and ~-50--75 mm March rainfall anomalies.

Area SE of Nairobi has had +6-7°C LST Anomalies, and ~-50--75 mm March rainfall anomalies.

Number of Rain Days

in past 30 days, as of 17 Apr 2011

Nr. Rain Days Anomaly
Current – Average of 2001–2009
**Setting**

- **Data pairs** of two attributes $X$ & $Y$, measured at $N$ sampling units:

  \[
  x = \begin{bmatrix}
  x_1 \\
  \vdots \\
  x_n \\
  \vdots \\
  x_N
  \end{bmatrix}\quad \text{and} \quad y = \begin{bmatrix}
  y_1 \\
  \vdots \\
  y_n \\
  \vdots \\
  y_N
  \end{bmatrix}
  \]

  - there are $N$ pairs of attribute values $\{(x_n, y_n), \quad n = 1, \ldots , N\}$

- **Scatter plot**: graph of $y$- versus $x$-values in attribute space: $y$-values serve as coordinates in vertical axis, $x$-values as coordinates in horizontal axis; $n$-th point in scatter-plot has coordinates $(x_n, y_n)$

- **Objective**: provide a quantitative *summary* of the above scatter plot as a measure of *association* between $x$- and $y$-values
Scatter Plot Quadrants

**Scatter plot center**: point \((\bar{x}, \bar{y})\) with coordinates equal to the data means: 
\[
\bar{x} = N^{-1} \sum_{n=1}^{N} x_n, \quad \bar{y} = N^{-1} \sum_{n=1}^{N} y_n
\]

**Scatter plot quadrants**: The line extending from the mean-x parallel to the y-axis and the line extending from the mean-y parallel to the x-axis define 4 quadrants in the scatter-plot.

**Deviations from the mean**: any measure association between \(X\) and \(Y\) should be independent of where the sample scatter plot is “centered”. Consequently, we’ll be looking at deviations of the data from their respective means: \((x_n - \bar{x}, y_n - \bar{y})\)

- quadrant I: \(x_n - \bar{x} > 0, y_n - \bar{y} > 0\)
- quadrant II: \(x_n - \bar{x} < 0, y_n - \bar{y} > 0\)
- quadrant III: \(x_n - \bar{x} < 0, y_n - \bar{y} < 0\)
- quadrant IV: \(x_n - \bar{x} > 0, y_n - \bar{y} < 0\)
Since we are after a measure of association, we compute products of data deviations from their means. A large positive product indicates high $x$- and $y$-values of the same sign. A large negative product indicates high $x$- and $y$-values of different sign.

**Product signs in different quadrants:**

- **quadrant I:** $x_n - \bar{x} > 0 \& y_n - \bar{y} > 0 \therefore (x_n - \bar{x})(y_n - \bar{y}) > 0$
- **quadrant II:** $x_n - \bar{x} < 0 \& y_n - \bar{y} > 0 \therefore (x_n - \bar{x})(y_n - \bar{y}) < 0$
- **quadrant III:** $x_n - \bar{x} < 0 \& y_n - \bar{y} < 0 \therefore (x_n - \bar{x})(y_n - \bar{y}) > 0$
- **quadrant IV:** $x_n - \bar{x} > 0 \& y_n - \bar{y} < 0 \therefore (x_n - \bar{x})(y_n - \bar{y}) < 0$
Sample Covariance of a Scatter Plot

Products of deviations from means:

\[
\begin{bmatrix}
    x_1 - \bar{x} \\
    \vdots \\
    x_n - \bar{x} \\
    \vdots \\
    x_N - \bar{x}
\end{bmatrix} \odot \begin{bmatrix}
    y_1 - \bar{y} \\
    \vdots \\
    y_n - \bar{y} \\
    \vdots \\
    y_N - \bar{y}
\end{bmatrix} = \begin{bmatrix}
    (x_1 - \bar{x})(y_1 - \bar{y}) \\
    \vdots \\
    (x_n - \bar{x})(y_n - \bar{y}) \\
    \vdots \\
    (x_N - \bar{x})(y_N - \bar{y})
\end{bmatrix}
\]

where \(\odot\) denotes element-by-element multiplication of two arrays; there are \(N\) such products \(\{(x_n - \bar{x})(y_n - \bar{y}), n = 1, \ldots, N\}\)

Average of \(N\) products:

\[
\frac{1}{N} \sum_{n=1}^{N} (x_n - \bar{x})(y_n - \bar{y}) = \frac{1}{N} [x - \bar{x}1]^T [y - \bar{y}1]
\]

1 denotes a \((N \times 1)\) vector of 1s; superscript \(T\) denotes transposition

Sample covariance between data of attributes \(x\) and \(y\):

\[
\hat{\sigma}_{XY} = \frac{1}{N-1} \sum_{n=1}^{N} (x_n - \bar{x})(y_n - \bar{y})
\]

average of products of data deviations from their means; a measure of joint variability (association) between two attributes \(X\) and \(Y\)

Sample variance = covariance of an attribute with itself: \(x_n \equiv y_n\)

\[
x_n \equiv y_n \rightarrow \hat{\sigma}_{XX} = \hat{\sigma}_X = \frac{1}{N-1} \sum_{n=1}^{N} (x_n - \bar{x})^2
\]
Sample covariance between data of attributes $X$ and $Y$:

$$\hat{\sigma}_{XY} = \frac{1}{N - 1} \sum_{n=1}^{N} (x_n - \bar{x})(y_n - \bar{y})$$

sum of products of data deviations from their means, divided by $N - 1$

**Interpretation:**

- Large positive covariance indicates data pairs predominantly lying in quadrants I and III
- Large negative covariance indicates data pairs predominantly lying in quadrants II and IV
- Small covariance indicates data pairs lying in all quadrants, in which case positive and negative products cancel out when one computes their mean

**NOTE:** The covariance is a measure of linear association between $X$ and $Y$, and just a summary measure of the actual scatter plot
Sample Covariance and Correlation Coefficient

- **Problems with sample covariance:**
  - not easily interpretable, since $x$- and $y$-values can have different units and sample variances
    \[
    \hat{\sigma}_X = \frac{1}{N-1} \sum_{n=1}^{N} (x_n - \bar{x})^2 \quad \hat{\sigma}_Y = \frac{1}{N-1} \sum_{n=1}^{N} (y_n - \bar{y})^2
    \]
  - sensitive to outliers; quantifies only *linear* relationships

- **Sample correlation coefficient:**
  - Pearson’s product moment correlation:
    \[
    \hat{\rho}_{XY} = \frac{\hat{\sigma}_{XY}}{\sqrt{\hat{\sigma}_X \hat{\sigma}_Y}} = \frac{[1/(N-1)] \sum_{n=1}^{N} (x_n - \bar{x})(y_n - \bar{y})}{\sqrt{[1/(N-1)] \sum_{n=1}^{N} (x_n - \bar{x})^2 \sqrt{[1/(N-1)] \sum_{n=1}^{N} (y_n - \bar{y})^2}}
    \]
  - lies in $[-1, +1]$; sensitive to outliers; quantifies only *linear* relationships

- **Sample rank correlation coefficient:** (Spearman’s correlation):
  - rank transform each sample data set, by assigning a rank of 1 to the smallest value and a rank of $N$ to the largest one
  - transform each data pair $\{x_n, y_n\}$ into a rank pair $\{r(x_n), r(y_n)\}$, where $r(x_n)$ and $r(y_n)$ is the rank of $x_n$ and $y_n$
  - compute the correlation coefficient of the rank pairs, as:
    \[
    \hat{\rho}_{XY}^S = 1 - \frac{6}{N(N^2 - 1)} \sum_{n=1}^{N} [r(x_n) - r(y_n)]^2
    \]
  - can detect non-linear monotonic relationships
Moment of Inertia of a Scatter Plot

**Motivation:** Instead of looking at average product of deviations from mean, we could look at the moment of inertia of a scatter plot; that is, the average squared distance between any pair \((x_n, y_n)\) and the 45° line; **Note:** such a line does not always make sense, but so be it for now.

\[
\cos(45) = \frac{d_n}{|x_n - y_n|} \Rightarrow d_n = \frac{\sqrt{2}}{2} |x_n - y_n| \Rightarrow d_n^2 = \frac{1}{2} (x_n - y_n)^2
\]

**Moment of inertia** = average deviation of scatter plot points from the 45° line:

\[
\hat{\gamma}_{XY} = \frac{1}{N} \sum_{n=1}^{N} d_n^2 = \frac{1}{2N} \sum_{n=1}^{N} (x_n - y_n)^2
\]

**Note:** The moment of inertia for a scatter plot aligned with the 45° line is always 0; that is, \(x_n = y_n\), \(\forall n\) \(\Rightarrow \hat{\gamma}_{xx} = \hat{\gamma}_{yy} = 0\)

alternatively, the dissimilarity of an attribute with itself is 0.
Recall:

\[
\hat{\sigma}_X = \frac{1}{N} \sum_{n=1}^{N} x_n^2 - \hat{\mu}_X^2
\]

\[
\hat{\sigma}_Y = \frac{1}{N} \sum_{n=1}^{N} y_n^2 - \hat{\mu}_Y^2
\]

\[
\hat{\sigma}_{XY} = \frac{1}{N} \sum_{n=1}^{N} x_n y_n - \hat{\mu}_X \hat{\mu}_Y
\]

Expanding:

\[
\hat{\gamma}_{XY} = \frac{1}{N} \sum_{n=1}^{N} (x_n - y_n)^2 = \frac{1}{N} \sum_{n=1}^{N} (x_n^2 + y_n^2 - 2x_n y_n)
\]

\[
= \frac{1}{N} \sum_{n=1}^{N} x_n^2 + \frac{1}{N} \sum_{n=1}^{N} y_n^2 - 2 \frac{1}{N} \sum_{n=1}^{N} x_n y_n
\]

\[
= \hat{\sigma}_X + \hat{\mu}_X^2 + \hat{\sigma}_Y + \hat{\mu}_Y^2 - 2 \hat{\sigma}_{XY} - 2 \hat{\mu}_X \hat{\mu}_Y
\]

\[
= \hat{\sigma}_X + \hat{\sigma}_Y - 2 \hat{\sigma}_{XY} + [\hat{\mu}_X - \hat{\mu}_Y]^2
\]

What’s the difference: To estimate the moment of inertia \(\gamma_{XY}\) you do not need to know the mean values \(\mu_X\) and \(\mu_Y\); these two mean values are required for estimating the covariance \(\sigma_{XY}\).
Geometric Interpretation (I)

- **Vector length:**
  \[ L_x = \sqrt{\sum_{n=1}^{N} x_n^2} \]
  - length = distance of point with coordinates \(\{x_1, \ldots, x_N\}\) from origin

- **Vector-scalar multiplication:** Multiplication of a vector by a scalar \(c\) changes length (and direction, depending on sign of \(c\)):
  \[ y = cx \Rightarrow L_y = |c| L_x \]
  - \(c > 1\) = expansion, \(c < 1\) = contraction; unit vector \(u = L_x^{-1} x\)

- **Inner product of two vectors:**
  \[ <x, y> = x^T y = \sum_{n=1}^{N} x_n y_n \]
  - a scalar quantity (could be negative, zero or positive)

- **Vector length:**
  \[ L_x = \sqrt{<x, x>} = \sqrt{x^T x} \]
  - inner product of a vector with itself

- **Angle \(\theta\) between two vectors \(y\) and \(x\):**
  \[ \cos(\theta) = \frac{<y, x>}{L_y L_x} = \frac{y^T x}{\sqrt{y^T y} \sqrt{x^T x}} \]
  \[ \cos(90) = \cos(270) = 0 \Rightarrow y^T x = 0 \]
  - the two vectors \(y\) and \(x\) are perpendicular, i.e., \(y \perp x\)
Geometric Interpretation (II)

- Let \( \tilde{x} = x - \bar{x} \) denote the vector of deviations (centered vector).

- **Variance:**
  \[
  \hat{\sigma}_x = \frac{1}{N} \sum_{n=1}^{N} (x_n - \bar{x})^2 = \frac{1}{N} \tilde{x}^T \tilde{x} = \frac{1}{N} L^2_{\tilde{x}}
  \]
  Variance proportional to squared vector length.

- **Covariance:**
  \[
  \hat{\sigma}_{YX} = \frac{1}{N} \sum_{n=1}^{N} (x_n - \bar{x})(y_n - \bar{y}) = \frac{1}{N} \tilde{y}^T \tilde{x} = \frac{1}{N} < \tilde{y}, \tilde{x} >
  \]
  Covariance proportional to inner product of centered vectors \( \tilde{y} \) and \( \tilde{x} \).

- **Correlation Coefficient**
  \[
  \hat{\rho}_{YX} = \frac{\hat{\sigma}_{YX}}{\sqrt{\hat{\sigma}_Y \hat{\sigma}_X}} = \frac{1}{N} < \tilde{y}, \tilde{x} > \sqrt{\frac{1}{N} L^2_{\tilde{y}} \sqrt{\frac{1}{N} L^2_{\tilde{x}}}} = \frac{< \tilde{y}, \tilde{x} >}{L_{\tilde{y}} L_{\tilde{x}}} = \cos(\theta)
  \]
  Cosine of angle \( \theta \) between two centered vectors \( \tilde{y} \) and \( \tilde{x} \) in a \( N \)-dimensional space.

- **Interpretation**
  - \( \hat{\rho}_{YX} = 0 \), vectors \( x \) and \( y \) are orthogonal → the two attributes are uncorrelated.
  - \( \hat{\rho}_{YX} = 1 \), vectors \( x \) and \( y \) are co-linear and lie along the same direction → perfectly positively correlated attributes.
  - \( \hat{\rho}_{YX} = -1 \), vectors \( x \) and \( y \) are co-linear but lie along opposite directions → perfectly negatively correlated attributes.
Geometric Interpretation (III)

- **Projection vector:** ("shadow") of vector \( y \) onto vector \( x \):
  
  A new vector along the direction of \( x \); \( L^{-1}_x x \) has unit length
  
  \[
P(y, x) = \frac{y^T x}{x^T x} x = \frac{\langle y, x \rangle}{L_x} \frac{1}{L_x} x
\]

- **Projection length:**
  
  \[
  |P(y, x)| = \frac{|\langle y, x \rangle|}{L_x} = L_y \frac{|\langle y, x \rangle|}{|L_y L_x|} = L_y |\cos(\theta)|
  \]

  When \( y \perp x \), \( |P(y, x)| = 0 \)

- **Unit Vector**
  
  Let \( 1 \) denote a \((N \times 1)\) vector of 1s, with length \( L_1 = \sqrt{N} \). The vector \( u = \frac{1}{\sqrt{N}} \cdot 1 \) has length \( L_u = 1 \) and forms equal angles with each of the \( N \) coordinate axes of a \( N \)-dimensional space

- **The sample mean vector**
  
  Vector \( \bar{y} = [\bar{y}_n, n = 1, \ldots, N]^T \) derived by projecting \( y \) onto the unit vector \( 1 \):
  
  \[
  \bar{y} = y^T \left( \frac{1}{\sqrt{N}} \cdot 1 \right) \frac{1}{\sqrt{N}} 1 = \frac{1}{N} \sum_{n=1}^{N} y_n = \bar{y} 1
  \]

  Sample mean = multiple of \( 1 \) required to yield the projection of \( y \) onto the unit vector deviation vector; Note that \( y - \bar{y} 1 \) is orthogonal to \( \bar{y} 1 \)

- **Regression = projection:**
  
  The regression of a centered vector \( \tilde{y} \) on another centered vector \( \tilde{x} \) is the projection of the former on the latter:
  
  \[
  \frac{\tilde{y}^T \tilde{x}}{\tilde{x}^T \tilde{x}} \tilde{x} = \frac{\sigma_Y X}{\sigma_X} \tilde{x} = \frac{\rho Y X \sqrt{\sigma_X} \sqrt{\sigma_Y}}{\sqrt{\sigma_X} \sqrt{\sigma_X}} \tilde{x} = \rho Y X \frac{\sqrt{\sigma_Y}}{\sqrt{\sigma_X}} \tilde{x} = \beta_X \tilde{x}
  \]

  \( \beta_X = \text{slope of variable } X \) for simple linear regression
Multivariate data set: $N$ measurements on $K$ attributes $\{X_1, \ldots, X_K\}$ made at $N$ sampling units and arranged in an $(N \times K)$ matrix $X$:

$X_{nk} = n$-th measurement for the $k$-th variable $X_k$

$n$-th row contains $K$ measurements of different attributes at a single sampling unit

$k$-th column contains $N$ measurements of a single attribute at all $N$ sampling units

Multivariate sample mean:

Conditional multivariate mean vector:

$(K \times 1)$ vector of mean values for all $K$ attributes, computed only from those rows of $X$ whose entries satisfy some condition (or query)
Computing Multivariate Sample Statistics (II)

Matrix of means:

\[
\bar{X} = \begin{bmatrix}
\bar{x}_1 & \cdots & \bar{x}_k & \cdots & \bar{x}_K \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
\bar{x}_1 & \cdots & \bar{x}_k & \cdots & \bar{x}_K \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
\bar{x}_1 & \cdots & \bar{x}_k & \cdots & \bar{x}_K \\
\end{bmatrix} = \frac{1}{N} \mathbf{1} \mathbf{1}^T \mathbf{X}
\]

\(\mathbf{1} \mathbf{x}^T = \) outer product of vectors \(\mathbf{1}\) and \(\mathbf{x}\); a \((N \times N)\) matrix;

\(\mathbf{1} \mathbf{1}^T = (N \times N)\) matrix of 1s

Matrix of deviations from means:

\[
\tilde{X} = \mathbf{X} - \bar{X} \text{ of size } (N \times K):
\]

\[
\begin{bmatrix}
x_{11} - \bar{x}_1 & \cdots & x_{1k} - \bar{x}_k & \cdots & x_{1K} - \bar{x}_K \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
x_{n1} - \bar{x}_1 & \cdots & x_{nk} - \bar{x}_k & \cdots & x_{nK} - \bar{x}_K \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
x_{N1} - \bar{x}_1 & \cdots & x_{Nk} - \bar{x}_k & \cdots & x_{NK} - \bar{x}_K \\
\end{bmatrix} = \mathbf{X} - \frac{1}{N} \mathbf{1} \mathbf{1}^T \mathbf{X}
\]

Sum of cross-products between variables \(X_k\) & \(X_{k'}\) (columns \(k, k'\)):

\[
\sum_{n=1}^{N} (x_{nk} - \bar{x}_k)(x_{nk'} - \bar{x}_{k'})
\]
Computing Multivariate Sample Statistics (III)

- **Matrix of squares and cross-products:** \( \tilde{X}^T \tilde{X} \) of size \((K \times K)\):
  \[
  \tilde{X}^T \tilde{X} = \left( X - \frac{1}{N} 11^T X \right)^T \left( X - \frac{1}{N} 11^T X \right) = X^T \left( I - \frac{1}{N} 11^T \right) X
  \]
  where \( I \) denotes the \((N \times N)\) identity matrix

- **Sample covariance matrix:** \( \hat{\Sigma}_X \) of size \((K \times K)\):
  \[
  \hat{\Sigma}_X = \frac{1}{N-1} (X - \bar{X})^T (X - \bar{X}) = \frac{1}{N-1} X^T \left( I - \frac{1}{N} 11^T \right) X
  \]

  **\(k,k'\)-th entry of covariance matrix:**
  \[
  \hat{\sigma}_{X_k X_{k'}} = \frac{1}{N-1} (X - \bar{X})_k^T (X - \bar{X})_{k'} = \frac{1}{N-1} \sum_{n=1}^{N} (x_k - \bar{x}_k)(x_{k'} - \bar{x}_{k'})
  \]
  \( A_{k .} = k\)-th row of matrix \( A \); \( A . k = k\)-th column on \( A \)

- **Note:** In the presence of missing values, one should compute all variance and covariance values only from those \( N < N \) rows of matrix \( X \) with no missing values. This ensures that the resulting covariance matrix \( \Sigma \) is a valid one.

- **Conditional covariance matrix:** \((K \times K)\) covariance matrix between all \( K^2 \) pairs of attributes, computed only from those rows of \( X \) whose entries satisfy some condition (or query)