

Visualizing Empirical Data

Gerald Farin, Dianne Hansford
Arizona State University

November 2, 2008

Standard Deviation

Data: $\mathbf{x} = [x_1, \dots, x_n]^T$

Assumption:

$$\frac{1}{n} \sum_{i=1}^n x_i = 0$$

Standard deviation (= length of vector):

$$\sigma(\mathbf{x}) = \sqrt{\frac{1}{n}(x_1^2 + \dots + x_n^2)}.$$

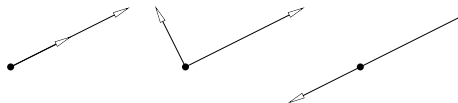
Correlation

$$\text{Data: } \mathbf{x} = [x_1, \dots, x_n]^T$$

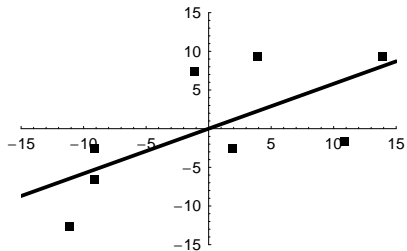
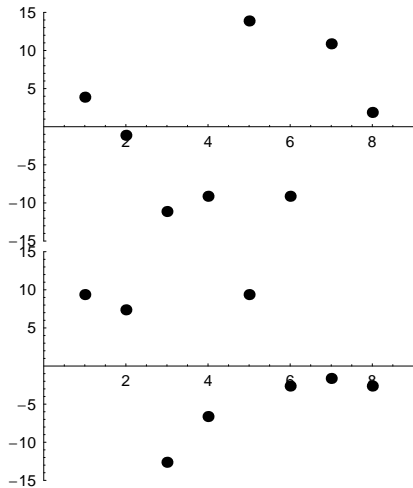
$$\mathbf{y} = [y_1, \dots, y_n]^T$$

Correlation $\rho \approx$ angle:

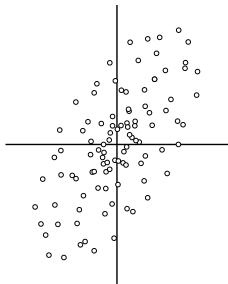
$$\rho = \cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}^T \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}.$$



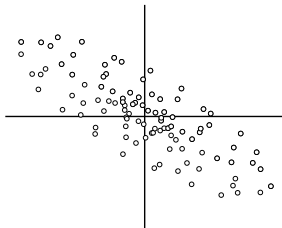
Scatter Plots



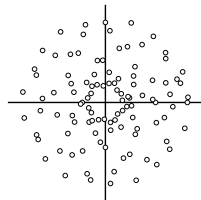
Correlation Examples



$$\rho = 1$$



$$\rho = -1$$



$$\rho = 0$$

Regression Line

Find a for $y = ax$ from

$$y_1 = ax_1$$

$$\vdots$$

$$y_n = ax_n.$$

Matrix form: $\mathbf{y} = \mathbf{x} \cdot a$.

Solution from

$$\mathbf{x}^T \mathbf{x} \cdot a = \mathbf{x}^T \mathbf{y}.$$

$\mathbf{x}^T \mathbf{x} = \text{scalar}$:

$$a = \frac{\mathbf{x}^T \mathbf{y}}{\mathbf{x}^T \mathbf{x}}$$

PCA

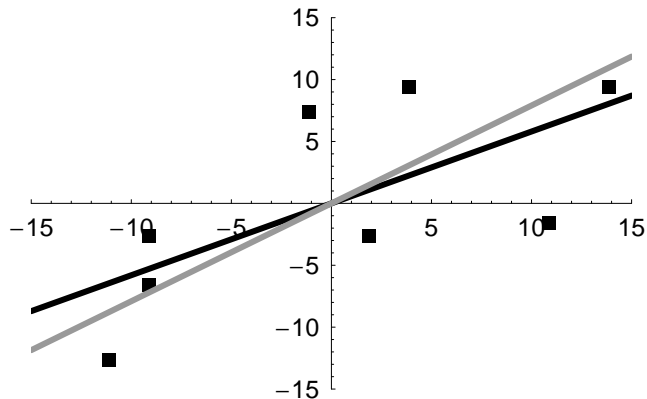
Data:

$$R = \begin{bmatrix} x_1 & y_1 \\ \vdots & \vdots \\ \vdots & \vdots \\ x_n & y_n \end{bmatrix}$$

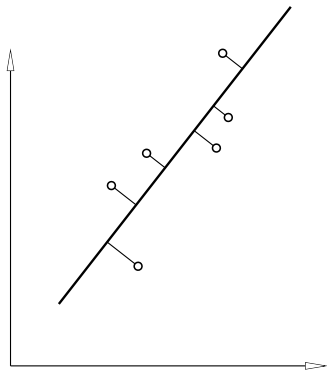
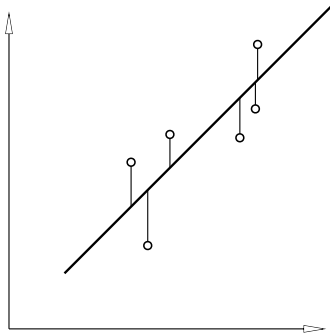
Shape of R : eigenvectors of $R^T R$

Dominant eigenvector \approx dominant line

Regression Line and Dominant Line



Regression Line and Dominant Line



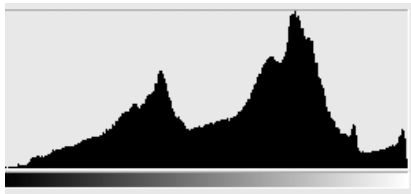
Histograms

A graphic representation of the distribution of tones within an image.

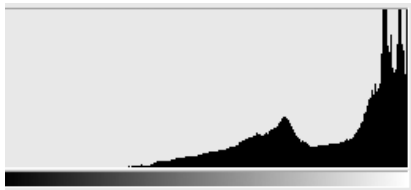
x-axis: possible color values (one bin per value).

y-axis: number of pixels in image having the x-value tone.

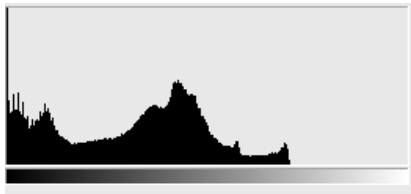
Histogram: normal



Histogram: overexposed

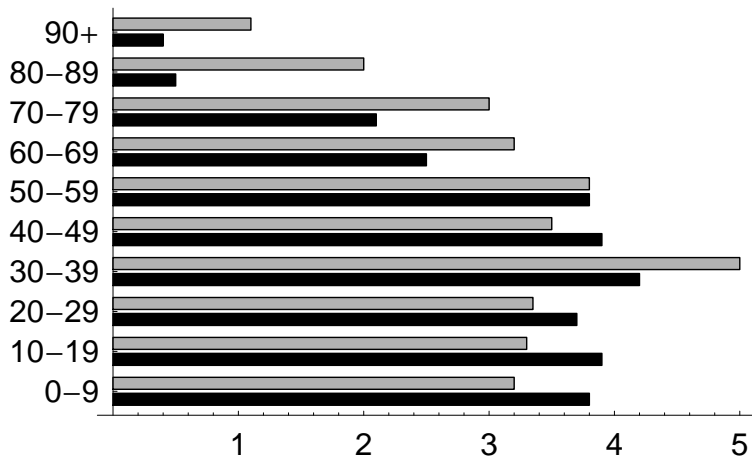


Histogram: underexposed



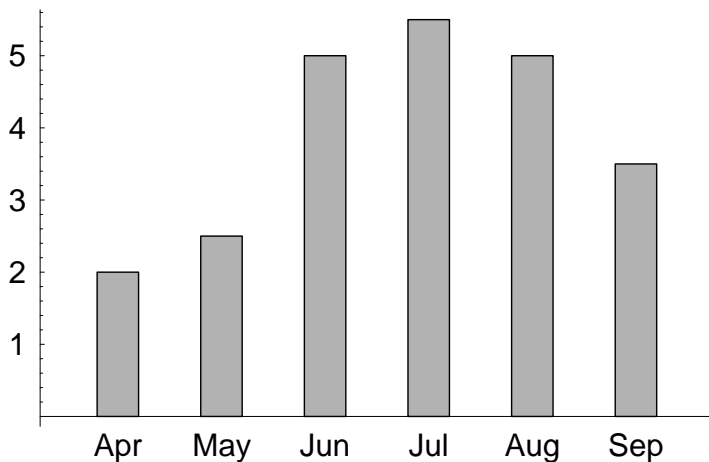
Population Histogram

2001 UK Population in Millions



Bar Charts

2008 Tourist Season Projected Profits in Millions

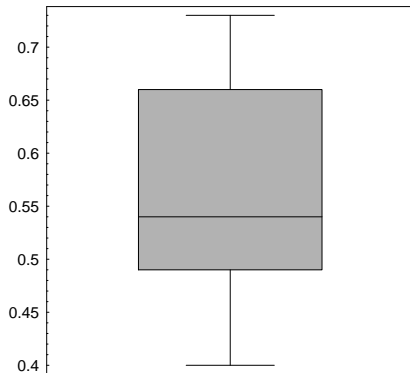
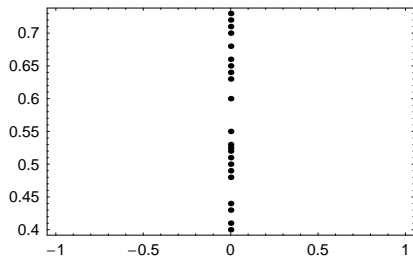


Comparison

Bar Charts: x-partition is prescribed

Histograms: x-partition (bins) must be created

Box Plots



Box Plots

1. Compute the median \rightarrow line
2. Compute median of lower half of data: lower quartile \rightarrow line
3. Compute the median of upper half of data: upper quartile \rightarrow line

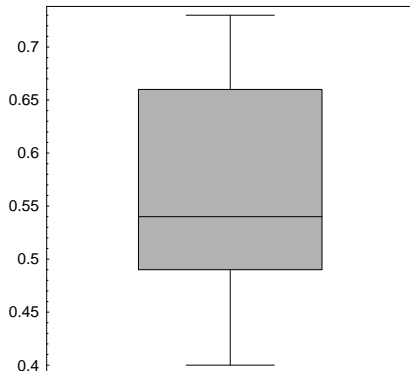
Outliers

Outliers:

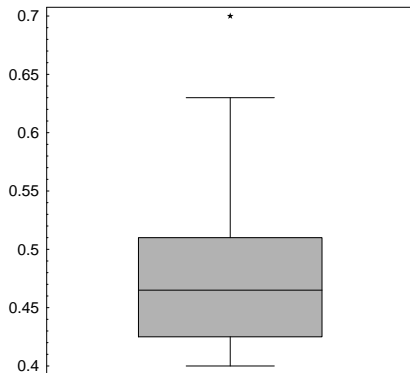
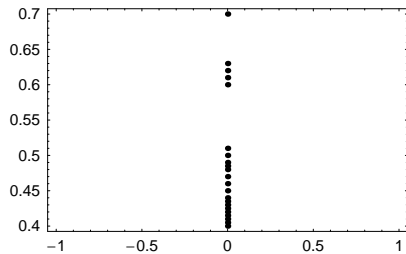
“way above” or “way below”
upper and lower quartiles

Whiskers:

largest and smallest non-outlier
data

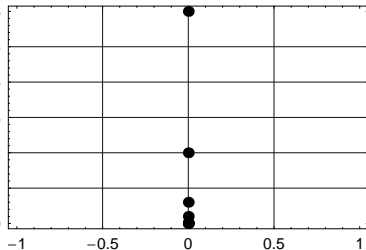


Biased Data

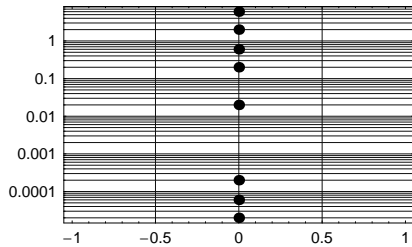
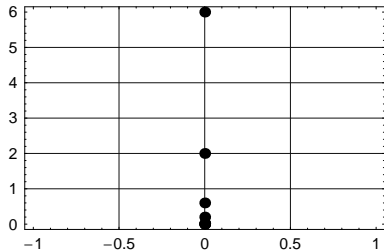


Large Variations

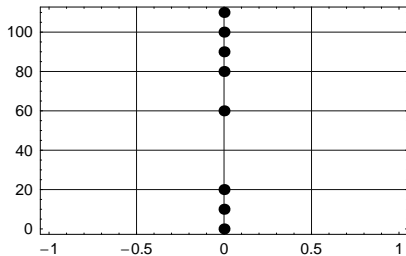
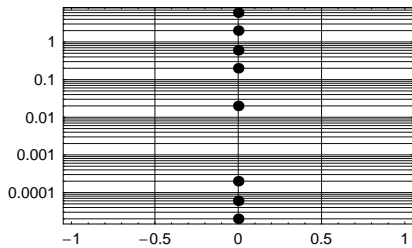
Source of sound	Pressure (Pa)
auditory threshold	0.00002
calm breathing	0.00006
very calm room	0.0002
normal talking	0.002
passenger car, 10m	0.02
jackhammer, 1m	2
jet engine, 100m	6



Log Plots



Example: Decibels



Example: Exponential Growth

