Exploratory Data Analysis (Wilks Ch. 3)

- Robustness
- Numerical Summaries
- Graphical Summaries
- Correlation
- Higher-Dimensional Data

Debra Baker

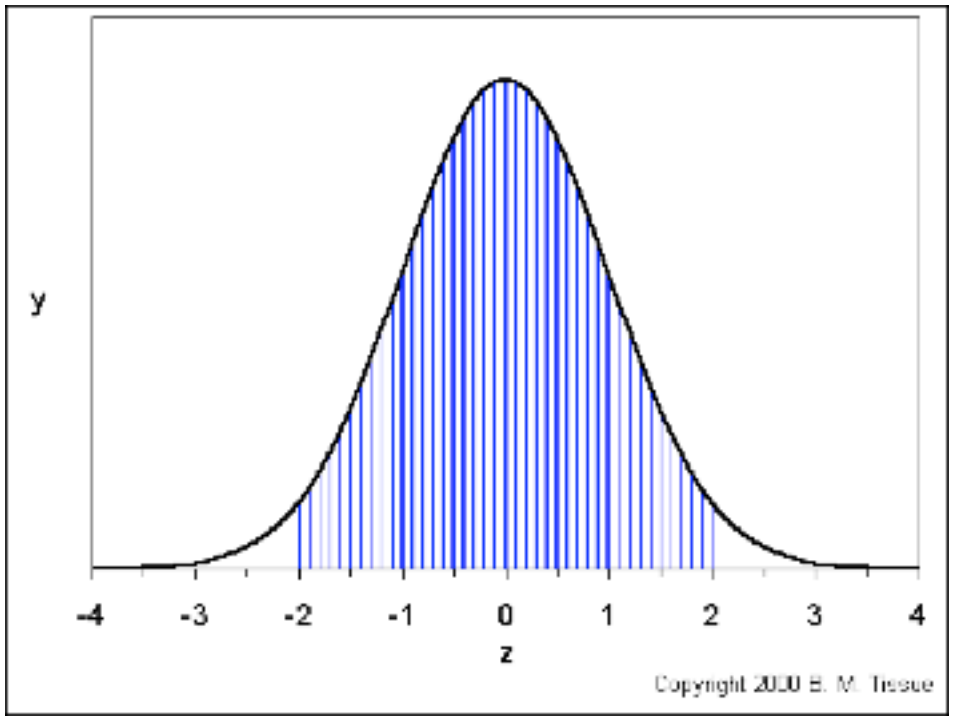
AOSC 630: Class #2

January 30, 2008

USGS gaging-station number	Gaging-station location	Period of record available for base-flow analysis	Average total flow (ft^3/s)	Average base flow (ft^3/s)	Ratio of base flow to total flow (in percent)
80765	Aplacles River at Apra Heights	10/1999 to 09/2001	nd	nd	nd
80850	Finale Creek at Agat	04/1960 to 12/1982	1.41	0.80	57
80940	Cetti River near Umatac	03/1960 to 09/1967	4.44	1.21	27
80950	La Sa Fua River near Agat	na	nd	nd	nd
80960	La Sa Fua River near Umatac	05/1953 to 06/1960, 10/1976 to 04/1984, 06/2000 to 09/2001	4.56	1.31	31
81000	Umatac River at Umatac	10/1952 to 12/1976	8.66	3.07	35
82100	Geus River near Merizo	05/1953 to 09/1975	3.08	0.72	24
83500	Inarajan River near Inarajan	10/1952 to 01/1983	17.44	6.61	38
84000	Timga River near Inarajan	11/1952 to 02/1986	5.62	2.03	36
84500	Talaeyuus River near Agat	10/1951 to 06/1960	21.06	7.21	34
84600	Talaeyuus River at mouth near Agat	06/1994 to 05/1995, 10/1996 to 03/1997, 07/1997 to 07/2002	nd	nd	nd
84700	Imong River near Agat	04/1960 to 03/1971, 10/1971 to 03/1994, 07/1997 to 09/2001	9.52	4.19	47
84800	Almagosa Springs near Agat	10/1951 to 12/1967, 12/1971 to 09/1975	3.03	0.83	26
84810	Almagosa River near Agat	04/1972 to 03/1992, 01/1995 to 04/1994, 03/1997 to 09/2001	5.98	2.05	36
84850	Maulap River near Agat	01/1972 to 03/1994, 07/1997 to 09/2000	5.14	2.18	42
84900	Fena Dam Spillway	10/1951 to 09/2001	na	na	na
85000	Talofofo River near Talofofo	12/1951 to 06/1962	50.95	15.38	30
85430	Ugum River above Talofofo Falls	05/1977 to 06/1993, 03/1997 to 09/2000	24.29	13.33	55
85500	Ugum River near Talofofo	06/1952 to 09/1970	29.44	15.48	53
85800	Vlig River near Yona	07/1952 to 03/1986, 04/1987 to 05/1995, 07/1997 to 09/2000	27.13	9.23	34
86200	Lonfit River near Ordot	10/1951 to 03/1960	10.42	2.80	27
86300	Pago River near	10/1951 to 09/1983, 11/1980 to 12/1980,	26.08	7.10	27

From: http://pubs.usgs.gov/wri/wri034126/htdocs/wrir03-4126.html

A good analysis method is insensitive to the assumptions about the data set.



From: http://www.chemicool.com/definition/gaussian_distribution.html

Common assumptions: "normal" distribution

Robust: performs reasonably well for most types of data

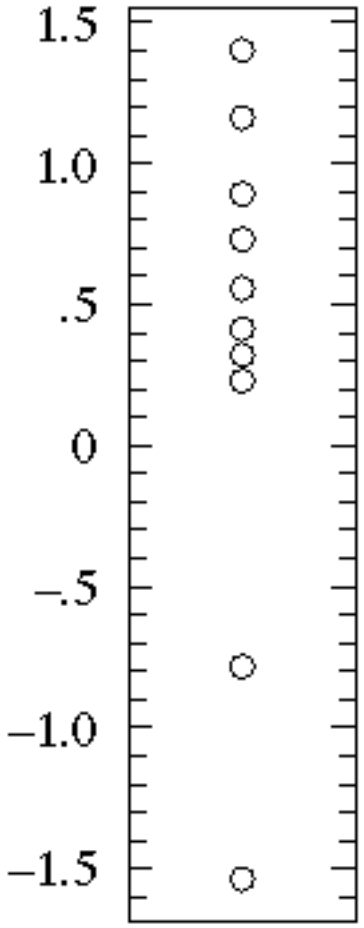
Resistant: not unduly influenced by a small number of outliers

There are three key features used to numerically describe a data set.

Location: the central tendency of the data set.

Spread: the dispersion of the data set around a central value.

Symmetry: how the data is distributed about the central value.



From: http://www.physics.csbsju.edu/stats/display.distribution.html

The first common numerical summary of a data set is a measure of its location.

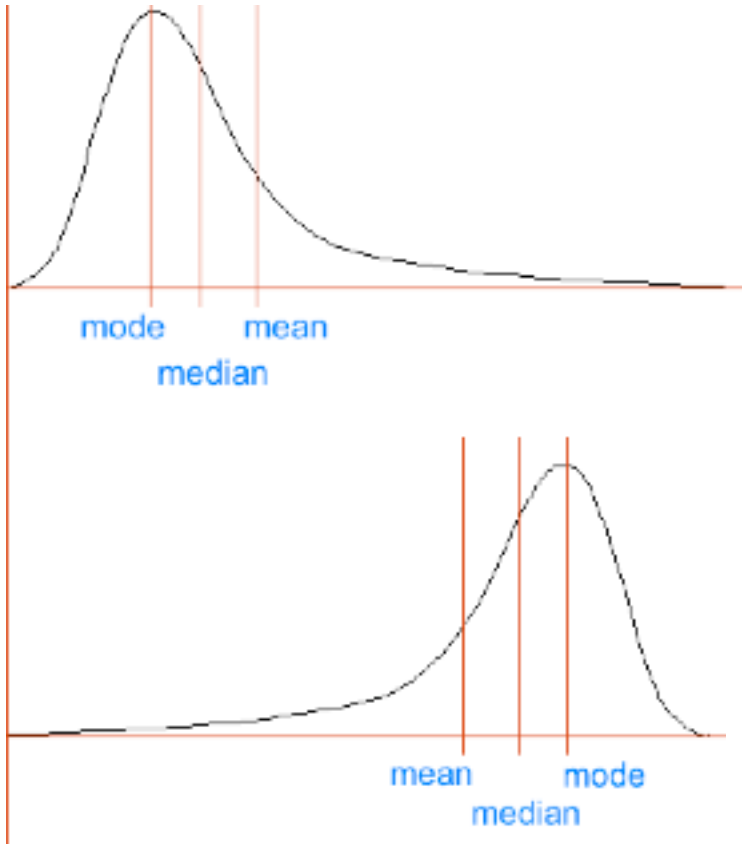
Mean: the average of all data points

Median: the center value in an ordered data set

Mode: the most frequently occurring value

Which of these measures are **robust?**

Which of these measures are **resistant?**



From: http://billkosloskymd.typepad.com/lexicillin_qd/2007/09/mean-vs-median-.html

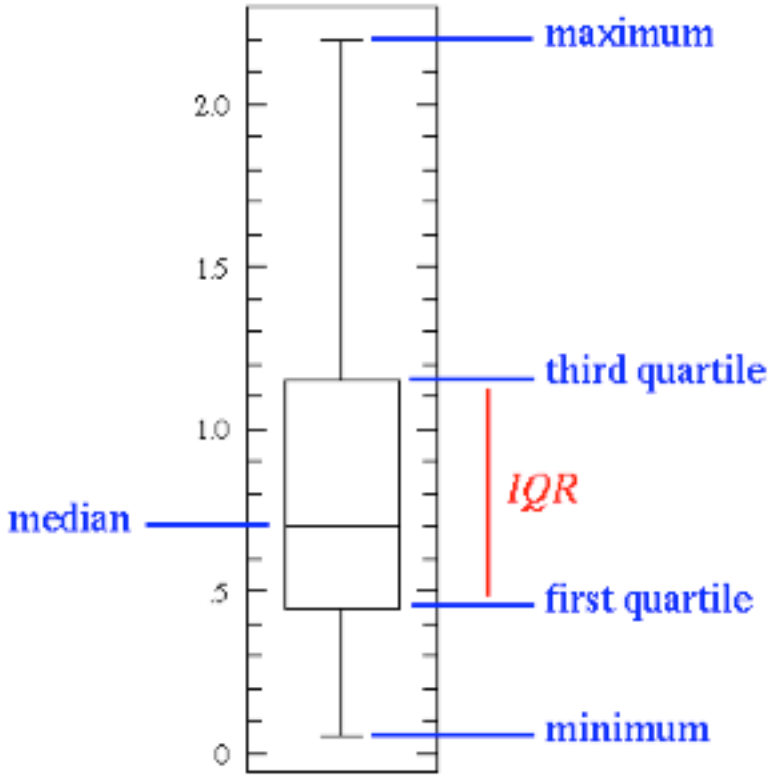
Quartiles divide the data set into four equal parts to describe its distribution.

First quartile: the middle of the data between the median and minimum.

Third quartile: the middle of the data between the median and maximum.

Are quartiles **robust** and **resistant**?

Quartiles are an example of a **quantile**,which can be based on any divisor (e.g., 10%).



From: http://www.physics.csbsju.edu/stats/display.distribution.html

The second common numerical summary of a data set is a measure of its spread.

Standard Deviation: the square root of the averaged square distance between data points and the mean.

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_1 - \bar{x}\right)^2}$$

Interquartile Range: specifies the range of the center 50% of the data.

$$IQR = q_{0.75} - q_{0.25}$$

Are these measures **robust**?

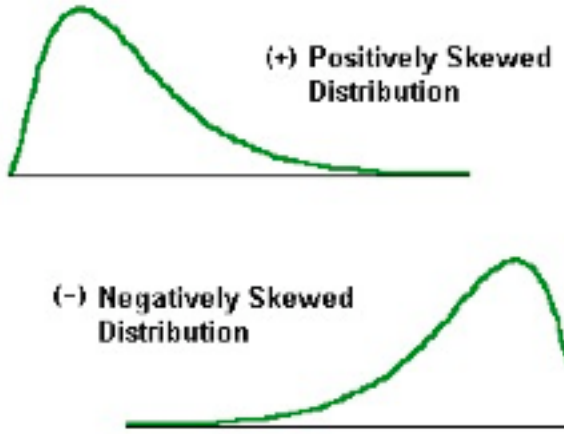
Are these measures **resistant**?

Equations 3.5 and 3.6 from Wilks (2006), pp. 26-27.

The third common numerical summary of a data set is a measure of its symmetry.

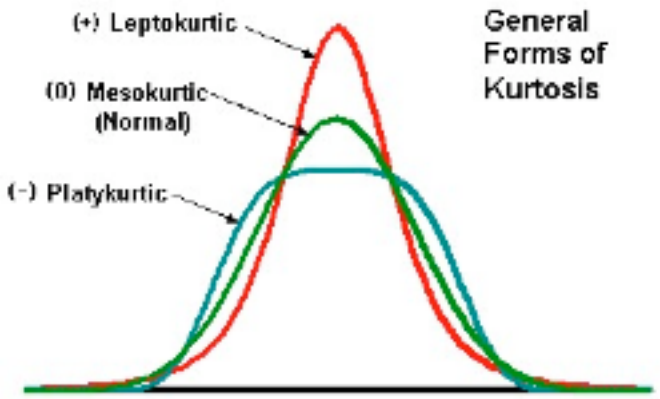
Positive Skewness:
distribution has a long right tail.

Negative Skewness:
distribution has a long left tail.



Positive Kurtosis: distribution has a tall narrow peak.

Negative Kurtosis: distribution has flat low peak.



From: http://mvpprograms.com/help/mvpstats/distributions/SkewnessKurtosis

There are two important measures of skewness.

Skewness Coefficient: a moments-based measure of symmetry.

$$\gamma = \frac{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^3}{s^3}$$

Yule-Kendall Index: compares the distance between the median and each of the two quartiles.

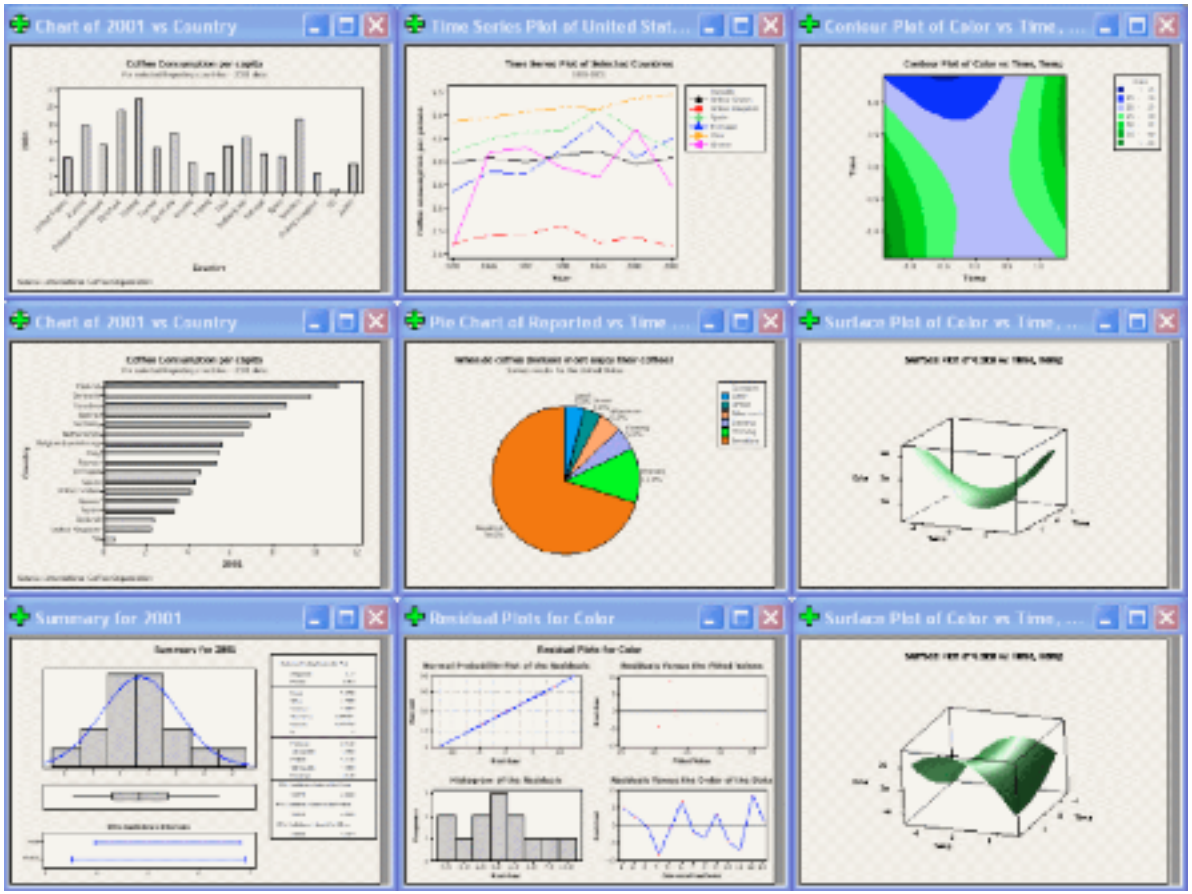
$$\gamma_{YK} = \frac{\left(q_{0.75} - q_{0.5}\right) - \left(q_{0.5} - q_{0.25}\right)}{IQR}$$

Are these measures **robust**?

Are these measures **resistant**?

Equations 3.9 and 3.10 from Wilks (2006), p. 28.

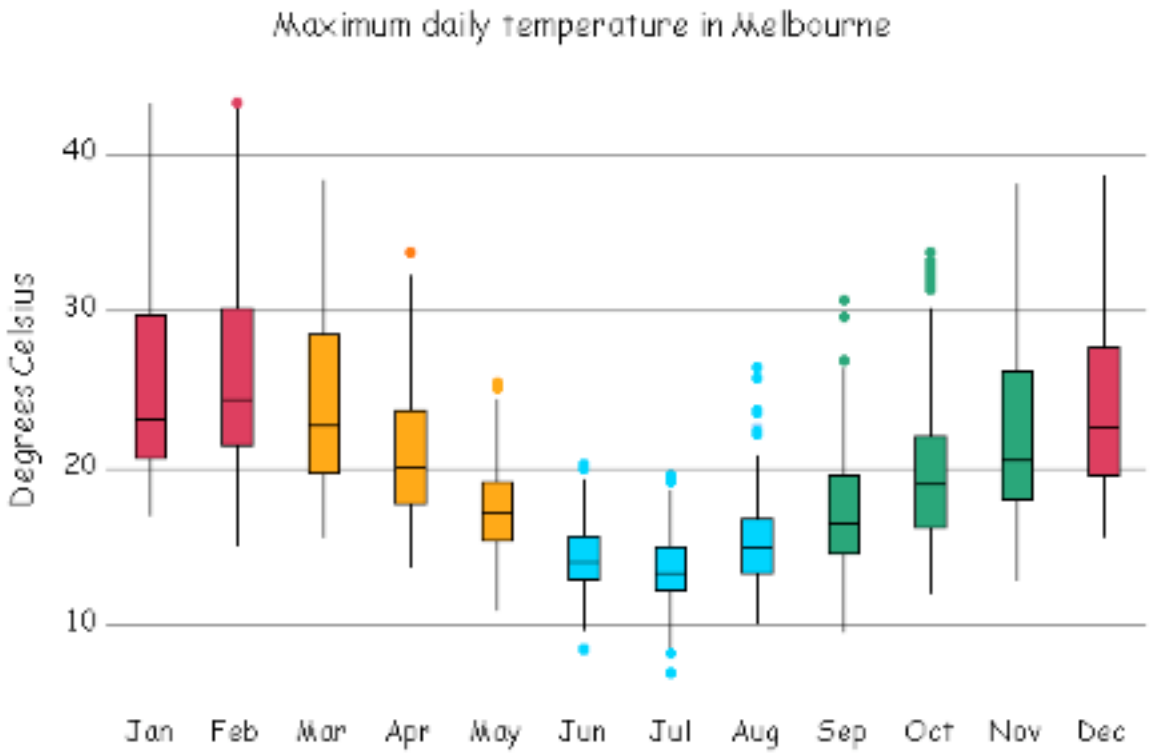
The characteristics of a data set are best conveyed using graphical summary techniques.



From: http://learnsigma.com/statistical-software-is-not-six-sigma/.

We will look specifically at **boxplots**, **histograms**, and **cumulative frequency distributions.**

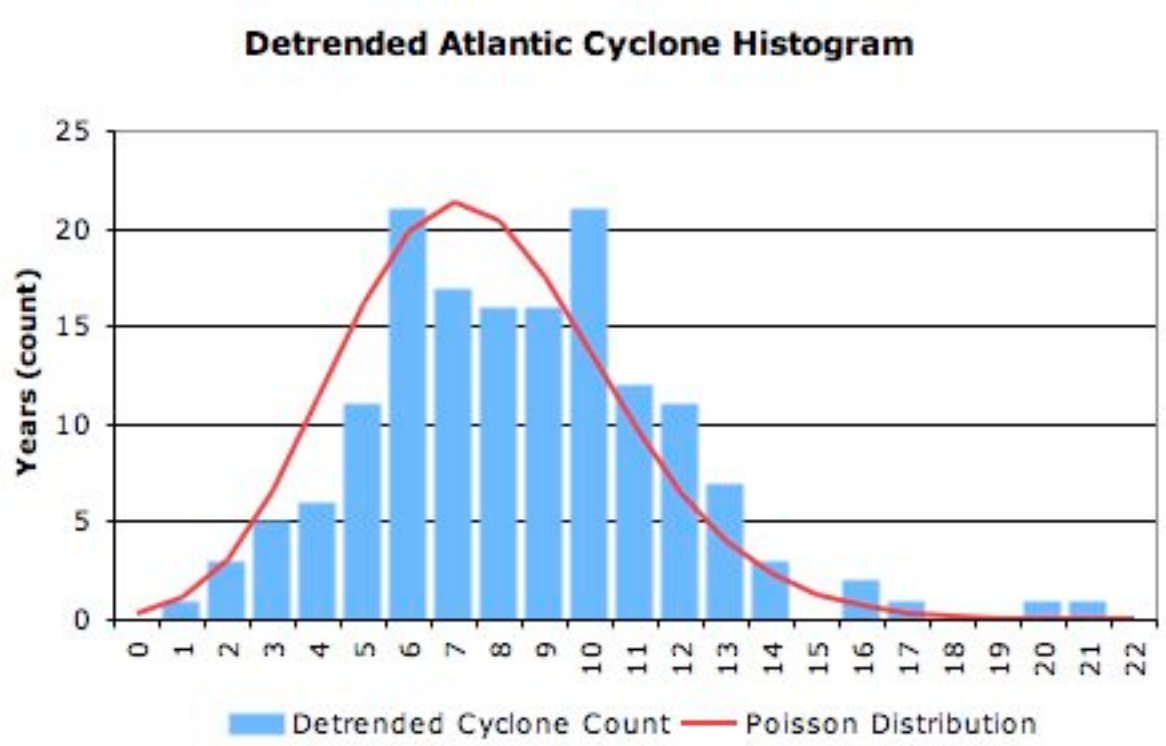
The boxplot is a simple graphic based on quartile calculations.



From:
http://www.scc.ms.unimelb.edu.au/whatisstatistics/weather.html

This type of boxplot is called a **schematic plot,** which highlights unusual extreme values.

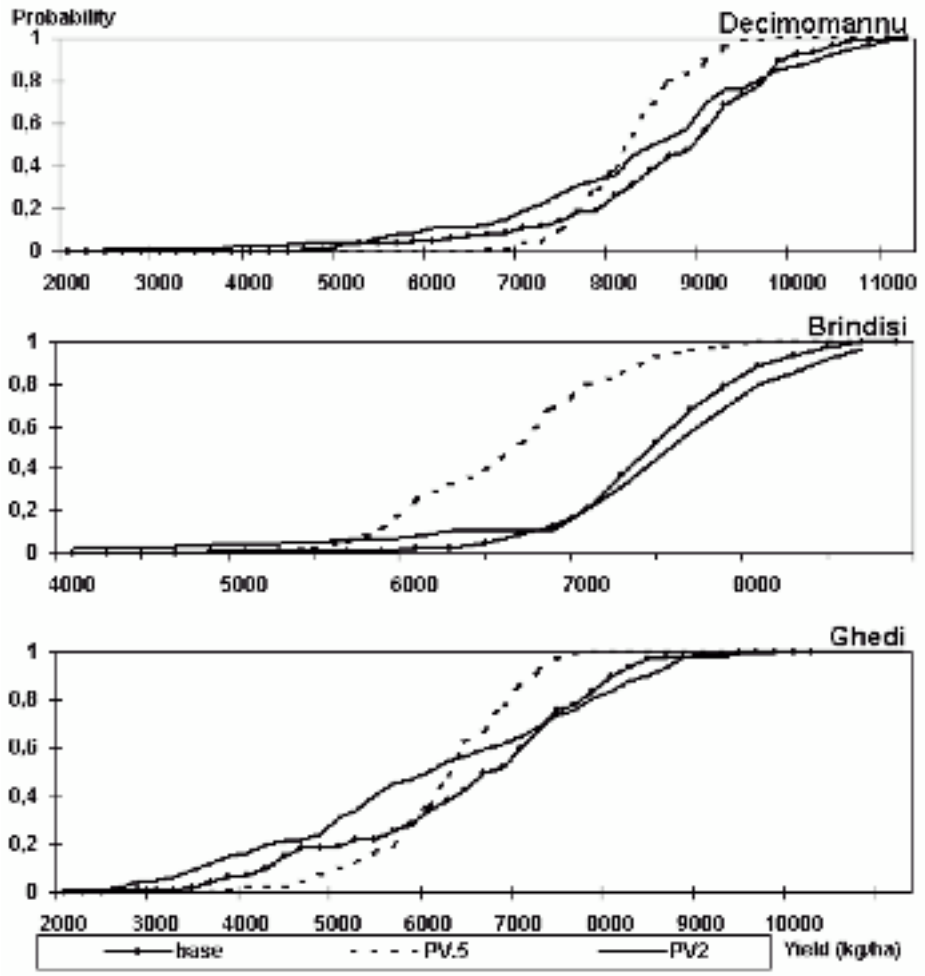
A histogram divides the data into bins and then charts the relative frequency distribution.



From:
http://www.climateaudit.org/?p=1022

Unlike a boxplot, a histogram can show if the data is **multimodal.**

The cumulative frequency distribution includes the number of points in all lower bins.



From: http://clima.casaccia.enea.it/staff/pona/Rep_Dec99/index.html

This graphic illustrates **wheat yield** from climate simulations changing the precipitation variance (PV)

For paired data sets, numerical and graphical summaries show the relationship of two variables.

Sample Paired Data

Set I		Set II	
x	y	x	y
0	0	2	8
1	3	3	4
2	6	4	9
3	8	5	2
5	11	6	5
7	13	7	6
9	14	8	3
12	15	9	1
16	16	10	7
20	16	20	17

From: Wilks (2006) p. 54

Scatterplots

Covariance

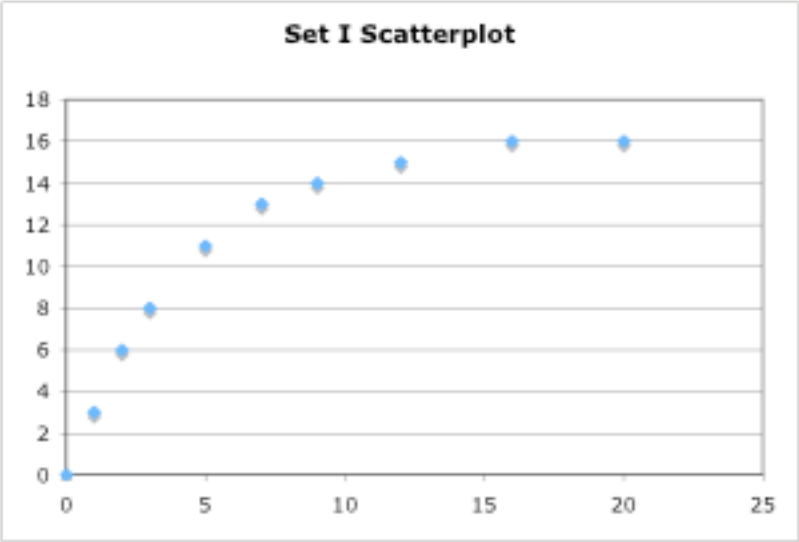
Pearson Correlation

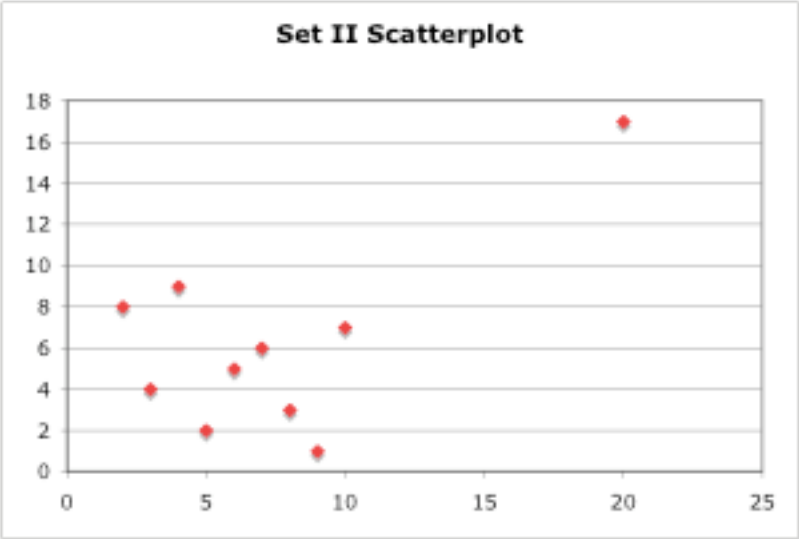
Rank Correlation

Lag Correlation

Autocorrelation

A scatterplot puts the paired data on an *x-y* grid.





Trends: Does the paired data show a recognizable pattern?

Curvature: Do the points show a positive or negative slope?

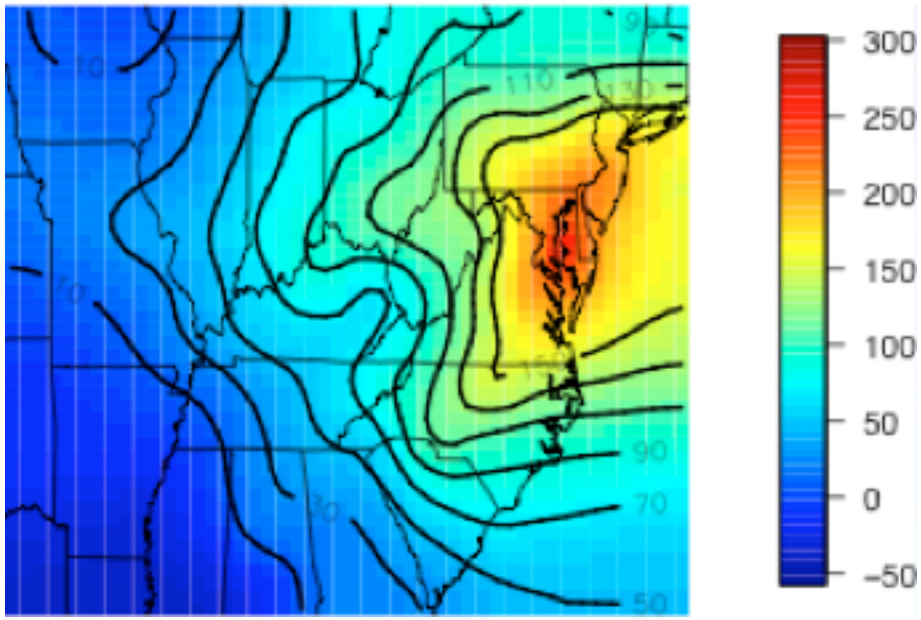
Clustering: How close together are the points?

Spread: What is the range of the data?

Outliers: Are there any unusually extreme values?

The covariance measures to what extent paired variables vary together.

$$Cov(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}\left[\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)\right]$$



From: http://www.image.ucar.edu/GSP/Projects/ResearchNuggets.shtml

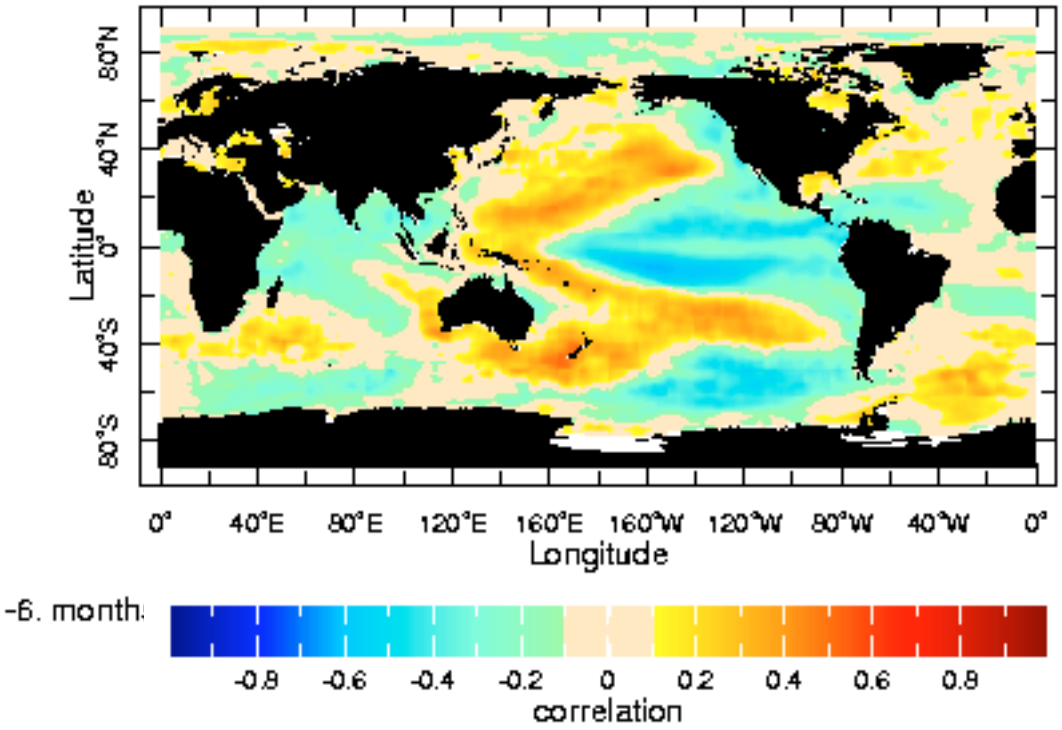
Cov > 0: If one variable rises so does the other.

Cov < 0: If one variable rises, the other falls.

This figure shows the estimated covariance of **surface ozone** in the Eastern U.S. between the grid point indicated by the red dot and the rest of grid points for one location.

The Pearson Correlation (i.e., standard correlation) is a product-moment coefficient of linear correlation.

$$r_{xy} = \frac{Cov(x,y)}{s_x s_y}$$



From: http://ingrid.ldeo.columbia.edu/dochelp/StatTutorial/Correlation/

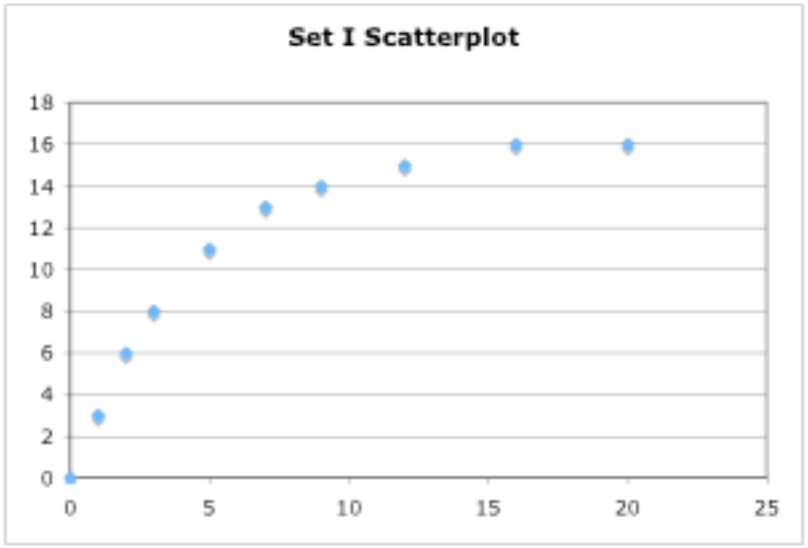
r = 1: perfect positive linear association
r = 0: no association
r = -1: perfect negative linear association

This figure shows the Pearson correlation of the **Southern Oscillation Index** and sea surface temperatures six months later.

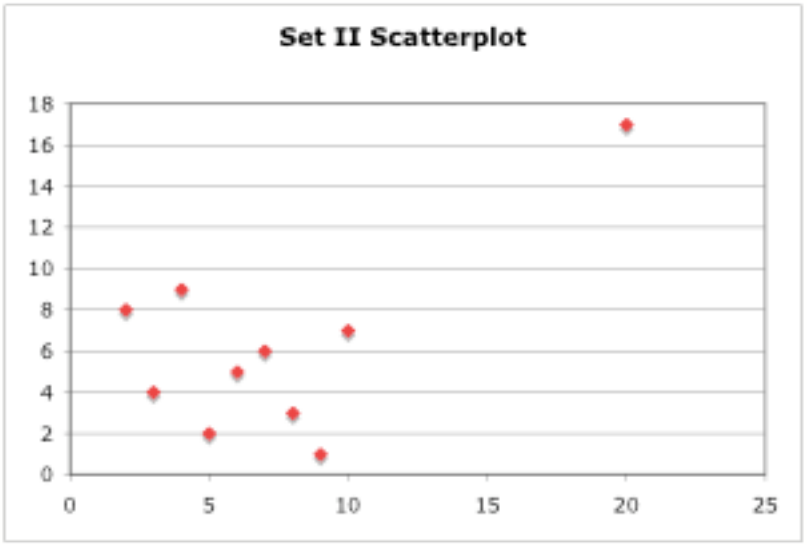
Is this measure **robust** and **resistant**?

What is the Pearson Correlation Coefficient for the sample paired data?

0.88



0.61



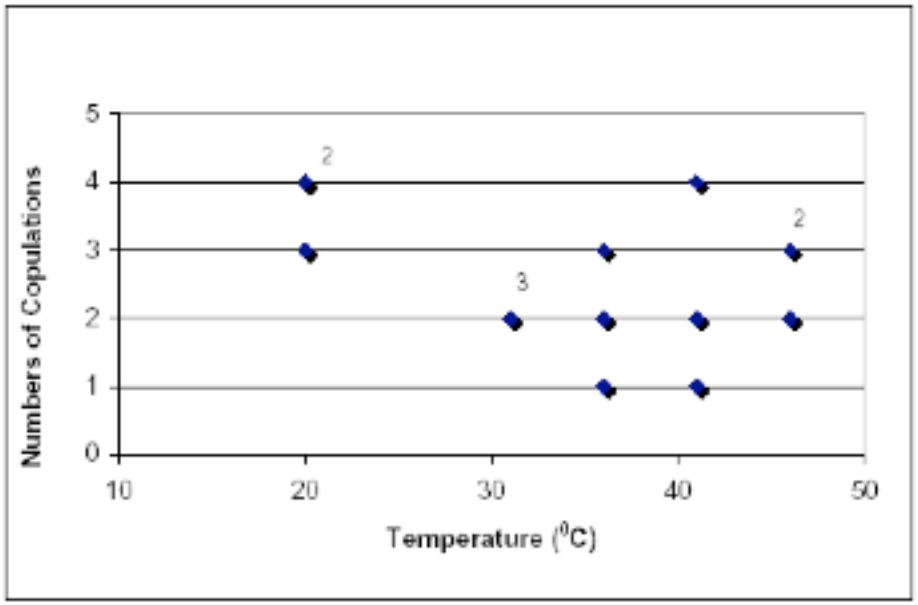
Will the correlation for each be **positive** or **negative**?

Which set will have the **higher magnitude** correlation?

Is the correlation meaningful?

The Spearman Rank Correlation is the Pearson correlation of the rank of the data, not its value.

$$r_{rank} = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n\left(n^2 - 1\right)}$$



From: Lauren Silbert's Penn State dissertation: The Aroid-Scarab Mutualism: Importance of Floral Temperature for Scarab Attraction and Copulation, available at http://www.pennscience.org/?view=past&year=2003&issue=1/2/&article=28

The *x* and *y* values are ordered **separately.**

The ordering is done from **smallest to largest.**

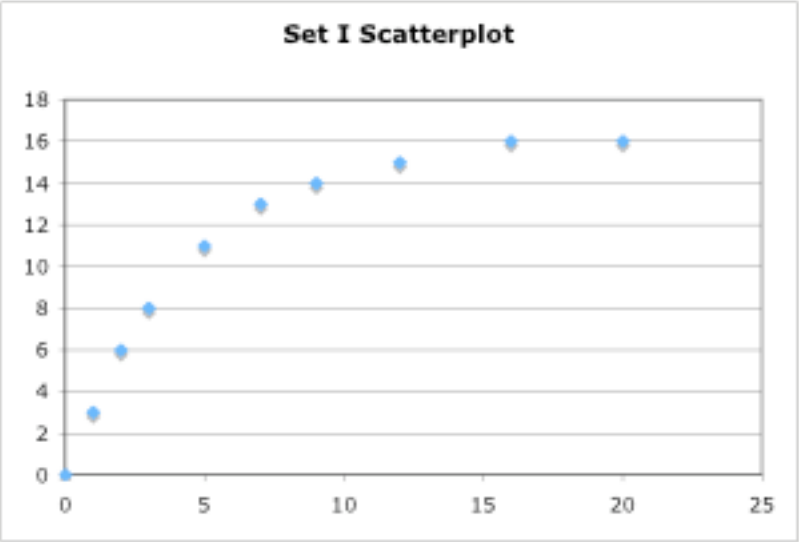
The value's rank number is then **substituted** for the value in the pairing.

Is this measure **robust** and **resistant**?

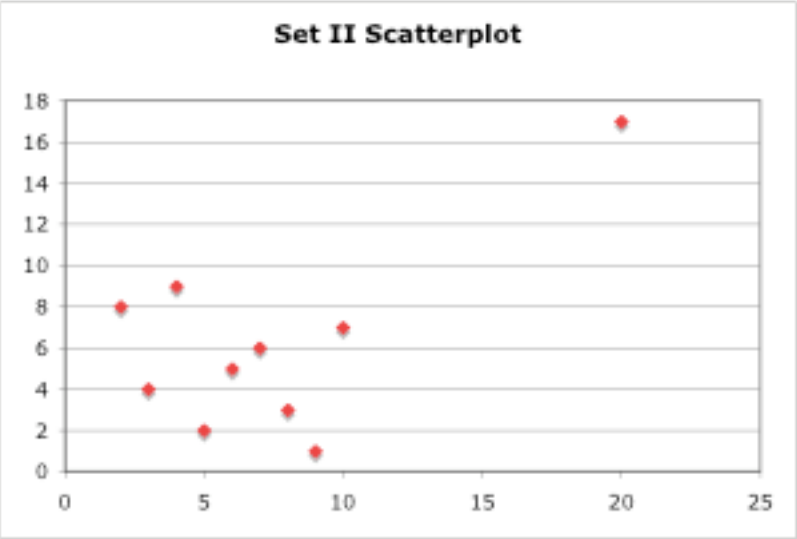
Does **beetle mating** correlate with temperature?
r_{xy} = 0.83 and r_{rank} = 0.11

What is the Spearman Rank Correlation Coefficient for the sample paired data?

1.00



0.018

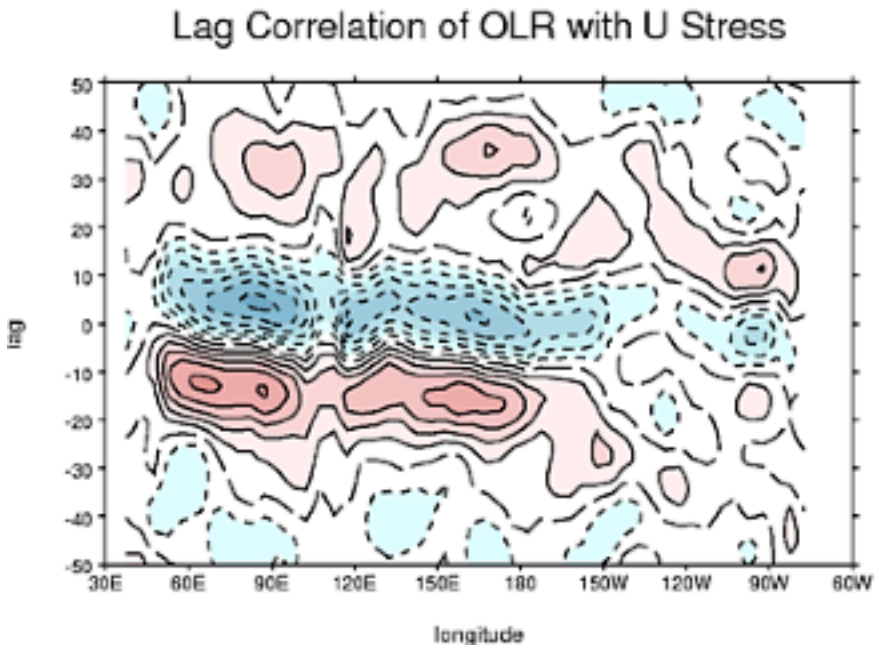


Will the rank correlation be **similar** to the Pearson correlation?

Since Set I has a **monotonic relationship** between *x* and *y*, what will its rank correlation be?

Lag correlation measures the association of *x* and *y* when they are separated by a time interval.

$$r_k = \rho\left(x_t, y_{t-k}\right)$$



Contours every 0.1, red indicates correlations > 0.1, blue indicates correlations < -0.1.

From: http://ugamp.nerc.ac.uk/hot/um/mjo.html

It is useful when one variable has a **delayed effect** on the other variable.

The **persistence** of a single variable over time can be measured by lag correlation.

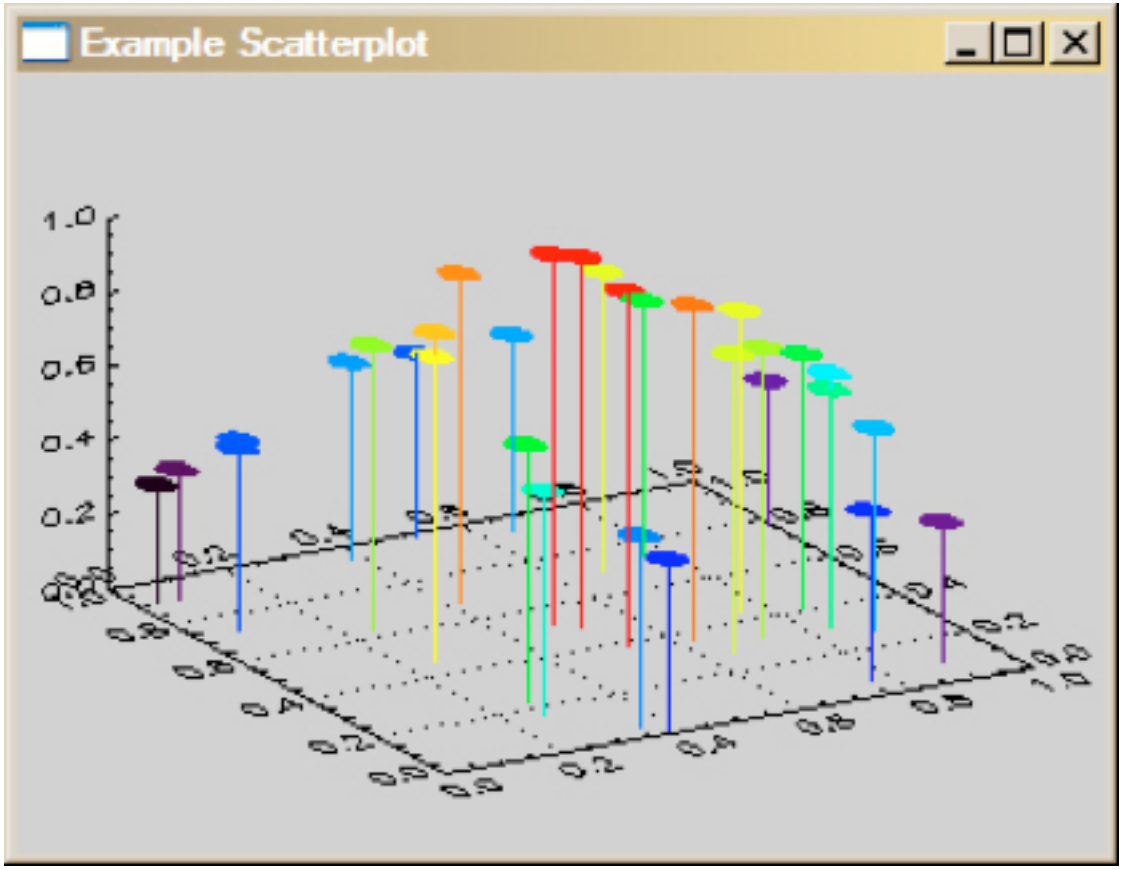
This relationship of a variable with itself separated by a time interval is called **autocorrelation**.

The relationship between multiple variables can be evaluated using more advanced techniques.

Covariance Matrix

Correlation Matrix

Correlation Maps



From: http://www.dfanning.com/tips/scatter3d.html

The covariance matrix is a collection of many covariances in a *d* x *d* matrix

The covariance matrix is always a **square matrix**.

It has one row/column for **each variable**.

The variance for each variable is on the **diagonal.**

The matrix is **symmetric** about the diagonal.

If all variables are of **unit** variance, then the covariance matrix is the same as the correlation matrix.

	x_1	x_2	x_3	x_4
x_1	v(1)	C(1,2)	C(1,3)	C(1,4)
x_2	C(2,1)	v(2)	C(2,3)	C(2,4)
x_3	C(3,1)	C(3,2)	v(3)	C(3,4)
x_4	C(4,1)	C(4,2)	C(4,3)	v(4)

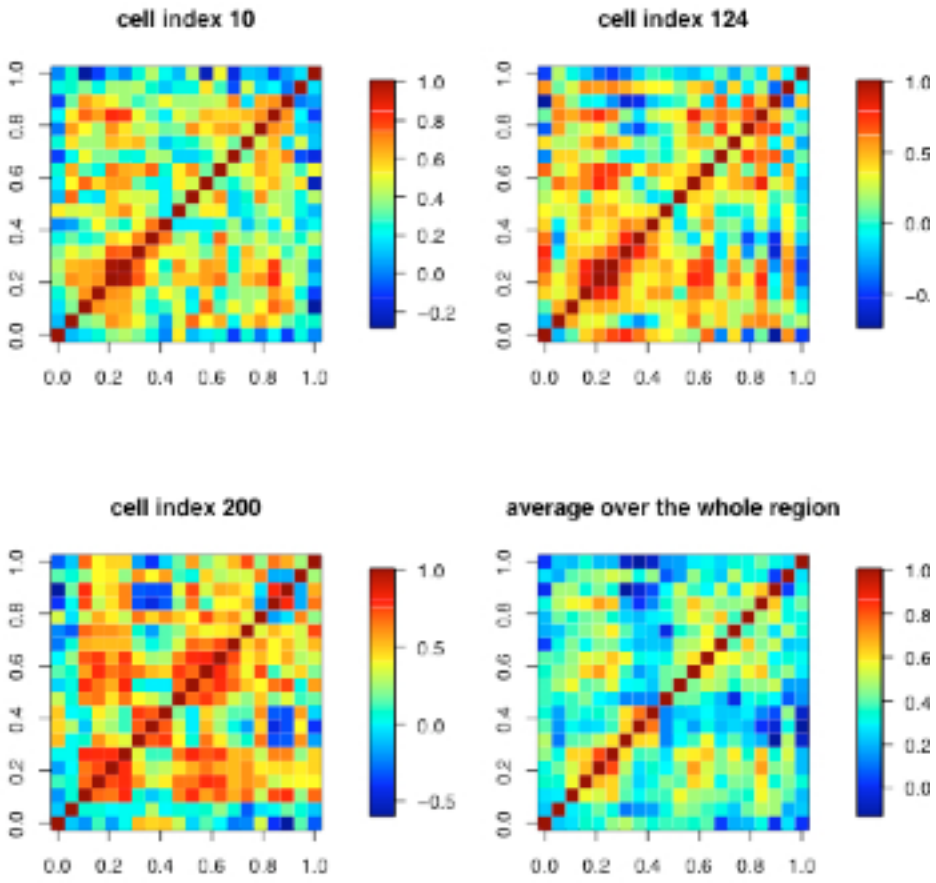
From: http://www.ucl.ac.uk/oncology/MicroCore/HTML_resource/Covariance_Matrix_popup.htm

A correlation matrix simultaneously displays correlations among matched multivariate data.

The main **diagonal** is the correlation of the variable with itself so it is always unity.

Many methods of multivariate analysis depend on it, such as **principal component analysis**.

This figure is an example of a correlation matrix among 20 **climate model biases** at certain spatial locations.



From: http://www.image.ucar.edu/GSP/Projects/ResearchNuggets.shtml

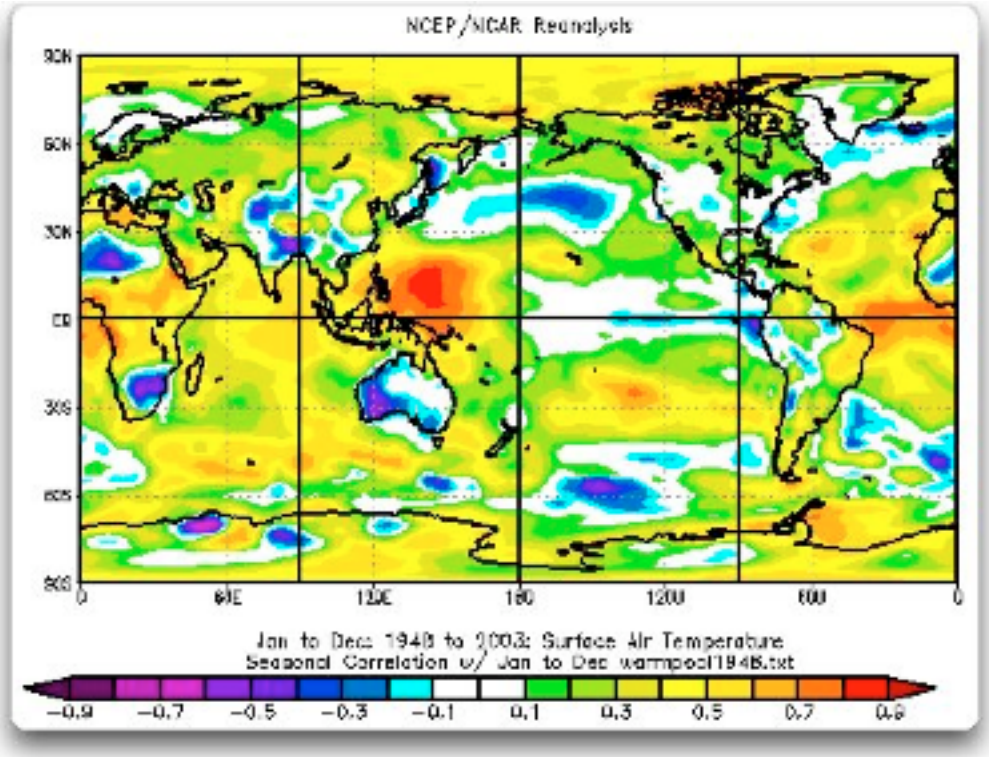
A correlation map arranges the correlation information by geographical location.

With large variable sets, a correlation matrix becomes **unwieldy.**

Atmospheric data is often too numerous when it is from a large number of **locations**.

A **one-point** correlation map shows the correlation between the variable at one location with the variable at all the other locations on the map.

Correlation between the sea surface temperature in the Pacific Warm Pool and surface air temperature at other locations



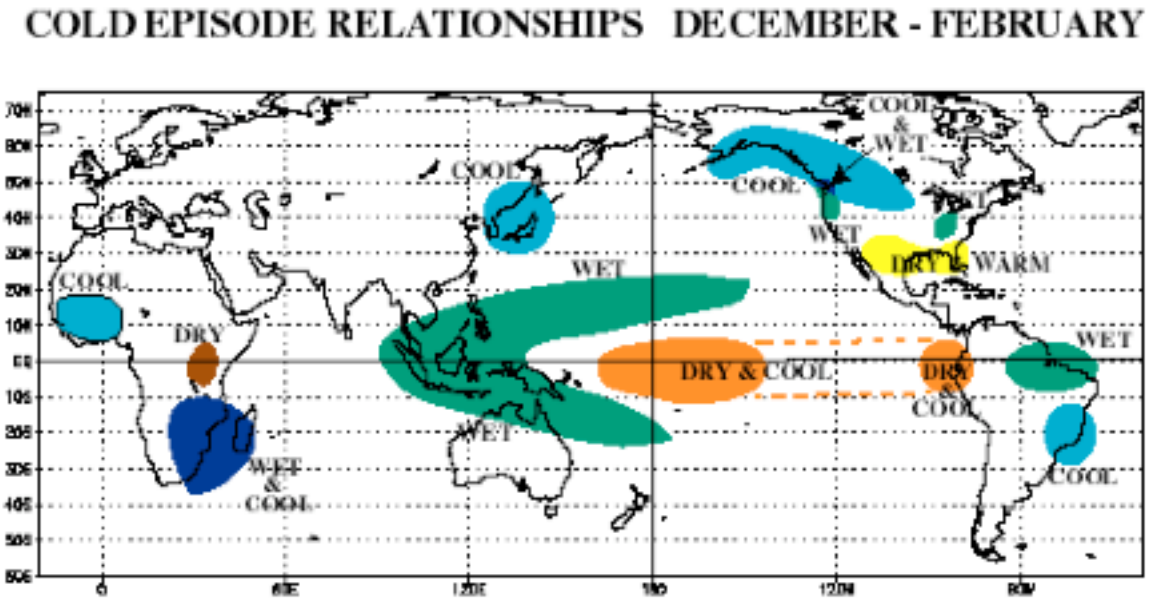
From: http://www.climateaudit.org/?p=858

Correlation maps are often used to detect teleconnections patterns.

A **teleconnection** is a strong statistical relationship between weather in different parts of the globe.

This figure shows the teleconnections of a **La Nina** event.

Recurring and persistent teleconnection patterns that spans vast areas are called **modes of variability.**



From: http://www.fas.usda.gov/pecad2/highlights/2000/04/eth/eth_drought.htm

In summary, Exploratory Data Analysis employs a variety of techniques to characterize data sets.

Robustness is the ability of an analysis method to provide reasonable results for different data sets.

Numerical Summaries include the mean, median, quartiles, standard deviation, interquartile range, skewness coefficient, and Yule-Kendall index.

Graphical Summaries include the boxplot, schematic plot, histograms, and cumulative frequency distributions.

Correlation measures the association between paired variables.

Higher-Dimensional Data can be analyzed using correlation and covariance matrices and maps.