Statistical Summary

ATM 305 – 12 November 2015

Review of Primary Statistics

- Mean

$$\overline{x} = \sum_{i=1}^{N} \frac{1}{N} x_i$$

x_i - scalar quantity
N - number of observations

- Median

Value at the 50% cumulative frequency distribution

Ex - 8, 25, 35, 60, **85**, 90, 115, 200, 825

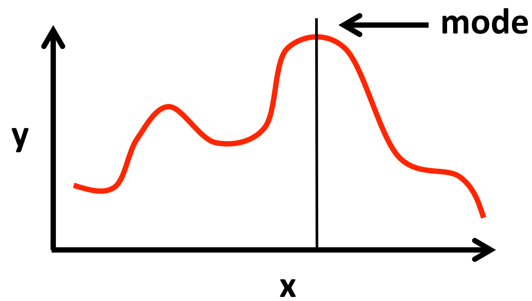
50% **50%**

- Mode

Maximum (or maxima) value of a distribution
Multiple modes possible (maxima)

$$y = f(x)$$

$$\frac{dy}{dx} = 0 \quad \frac{d^2y}{d^2x} < 0$$



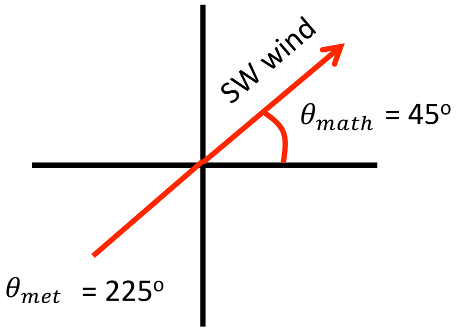
Wind

- Mean wind speed
  - Mean speed determined without consideration of direction of wind
- Mean resultant wind

$$\overline{\vec{V}} = \sum_{i=1}^{N} \frac{\vec{V}}{N}$$

$\vec{V}$ - vector wind

To calculate $\overline{\vec{V}}$ it is necessary to compute the individual u_i and v_i wind components, average them separately, and recombine $\bar{u}_i$ and $\bar{v}_i$ to get the mean vector wind.



Note the differences between mathematical wind direction and meteorological wind direction

Let θ = meteorological wind direction

$$\theta_{math} = 270° - \theta_{met}$$

$$\theta_{met} = 270° - \theta_{math}$$

Wind

- Mean resultant wind (cont.)

Using **mathematical** wind direction

$$u_i = \left|\vec{V}_i\right| \cos\theta_{math}$$

$$v_i = \left|\vec{V}_i\right| \sin\theta_{math}$$

Mean →

$$\bar{u}_i = \frac{1}{N}\sum_{i=1}^{N}\left|\vec{V}_i\right| \cos\theta_{math}$$

$$\bar{v}_i = \frac{1}{N}\sum_{i=1}^{N}\left|\vec{V}_i\right| \sin\theta_{math}$$

Wind

- Mean resultant wind (cont.)

Using **meteorological** wind direction

Note Changes

$$u_i = -\left|\vec{V}_i\right| \sin\theta_{met}$$

$$v_i = -\left|\vec{V}_i\right| \cos\theta_{met}$$

Mean →

$$\bar{u}_i = -\frac{1}{N}\sum_{i=1}^{N}\left|\vec{V}_i\right| \sin\theta_{met}$$

$$\bar{v}_i = -\frac{1}{N}\sum_{i=1}^{N}\left|\vec{V}_i\right| \cos\theta_{met}$$

Resultant Wind Speed

$$\left|\vec{V}_i\right| = \left(\bar{u}_i{}^2 + \bar{v}_i{}^2\right)^{\frac{1}{2}}$$

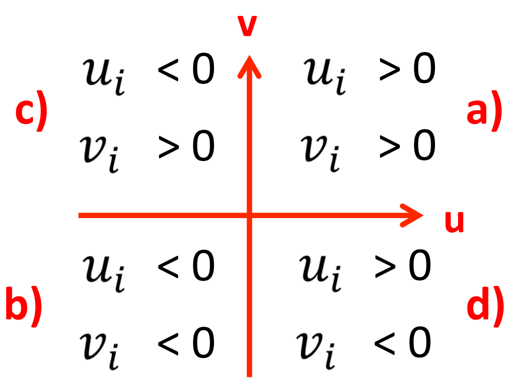
Resultant Wind Direction

$$\tan\bar{\theta} = \frac{\bar{v}_i}{\bar{u}_i} \rightarrow \bar{\theta} = \tan^{-1}\left(\frac{\bar{v}_i}{\bar{u}_i}\right)$$

Wind

- Mean resultant wind (cont.)

To get meteorological wind direction right, necessary to check individual coordinates



a)	$u_i > 0$	and	$v_i > 0$	(SW winds)	$180^o < \theta_{met} < 270^o$
b)	$u_i < 0$	and	$v_i < 0$	(NE winds)	$0^o < \theta_{met} < 90^o$
c)	$u_i > 0$	and	$v_i < 0$	(NW winds)	$270^o < \theta_{met} < 360^o$
d)	$u_i < 0$	and	$v_i > 0$	(SE winds)	$90^o < \theta_{met} < 180^o$

e)	$u_i = 0$	and	$v_i > 0$	(S winds)	$\theta_{met} = 180^o$
f)	$u_i = 0$	and	$v_i < 0$	(N winds)	$\theta_{met} = 0^o$
g)	$u_i > 0$	and	$v_i = 0$	(W winds)	$\theta_{met} = 270^o$
h)	$u_i < 0$	and	$v_i = 0$	(E winds)	$\theta_{met} = 90^o$

Wind

- Mean resultant wind (cont.)

Resultant wind speed is smaller than average speed of individual winds

The mean wind speed (NOT resultant wind) determines mean evaporation, cooling power, and boundary layer turbulence

Example: 10 wind measurements

5 northerly wind measurements, each at 10 kt
5 southerly wind measurements, each at 10 kt

What is the mean wind speed, and resultant wind speed?

mean wind speed = (10 + 10 + 10 + 10 + 10 + 10 + 10 + 10 + 10 + 10) / 10 = **10 kt**

resultant wind speed $\bar{u}_i$ = (0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0) / 10 = 0

$\bar{v}_i$ = (-10) + (-10) + (-10) + (-10) + (-10) + 10 + 10 + 10 + 10 + 10) / 10 = 0

$$\left|\vec{V}_i\right| = (\bar{u}_i{}^2 + \bar{v}_i{}^2)^{\frac{1}{2}} \quad = (0^2 + 0^2)^{\frac{1}{2}}$$ = **0 kt**

Wind

- Wind Persistence ("steadiness")

$$P = \frac{speed\ of\ resultant\ wind}{Mean\ wind\ speed}$$

P = 1 Winds always blows from the same direction (resultant wind and mean wind speed are equal)

P = 0 Wind is unsteady and blowing from all directions

Standard Deviation

Sometimes denoted by sigma, σ

$$s = \sigma = \left(\sum_{i=1}^{N} \frac{|x_i - \bar{x}|^2}{N} \right)^{\frac{1}{2}}$$

Individual Value → Mean → Anomaly

Note: $x_i = \bar{x} + x'$

$x' = x_i - \bar{x}$

Note: some texts when N is large replace N with (N – 1)

Variance:

$$s^2 = \sigma^2 = \sum_{i=1}^{N} \frac{|x_i - \bar{x}|^2}{N}$$

Simply the square of standard deviation

Both the standard deviation and variance measure the spread of a large sample of data

Standardized Anomalies

$$Z = \frac{x_i - \bar{x}}{\sigma}$$

Individual Value → x_i

Mean of that value → $\bar{x}$

Standard Deviation → σ

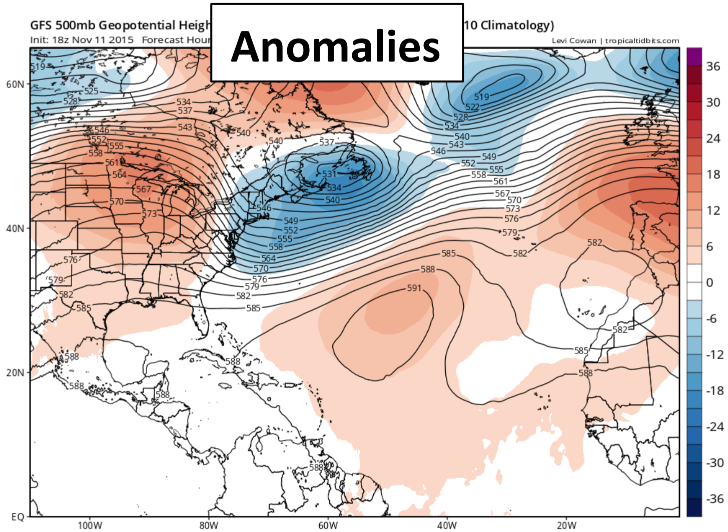
$$Z = \frac{x'}{\sigma}$$

Anomaly → x'

All that is needed is a mean and standard deviation and you can find the standardized anomaly of a variable

- Standardizing anomalies allows us to compare anomalous values in different parts of the globe
- Can be thought of as a measure of distance in standard deviation units of a value from its mean

Example: standardized anomalies of 500-hPa geopotential heights



See Link: http://www.tropicaltidbits.com/analysis/models/?model=gfs&region=atl&pkg=z500a_sd&

- Can now directly compare low heights over midwestern US to high heights over east Pacific
- Standard deviation of heights relatively low in tropics, but relatively high in mid-latitudes

Standardized Anomalies

Individual Value → x_i; Mean of that value → $\bar{x}$

$$Z = \frac{x_i - \bar{x}}{\sigma}$$

Standard Deviation → σ

$$Z = \frac{x'}{\sigma}$$

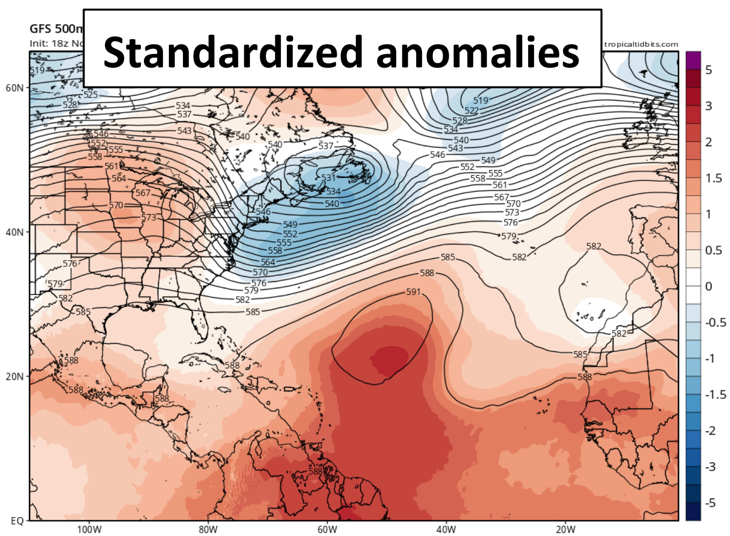
Anomaly → x'

All that is needed is a mean and standard deviation and you can find the standardized anomaly of a variable

Standardizing a dataset results in a **mean = 0** and a **standard deviation = 1**

- Standardizing anomalies allows us to compare anomalous values in different parts of the globe
- Can be thought of as a measure of distance in standard deviation units of a value from its mean

Example: standardized anomalies of 500-hPa geopotential heights



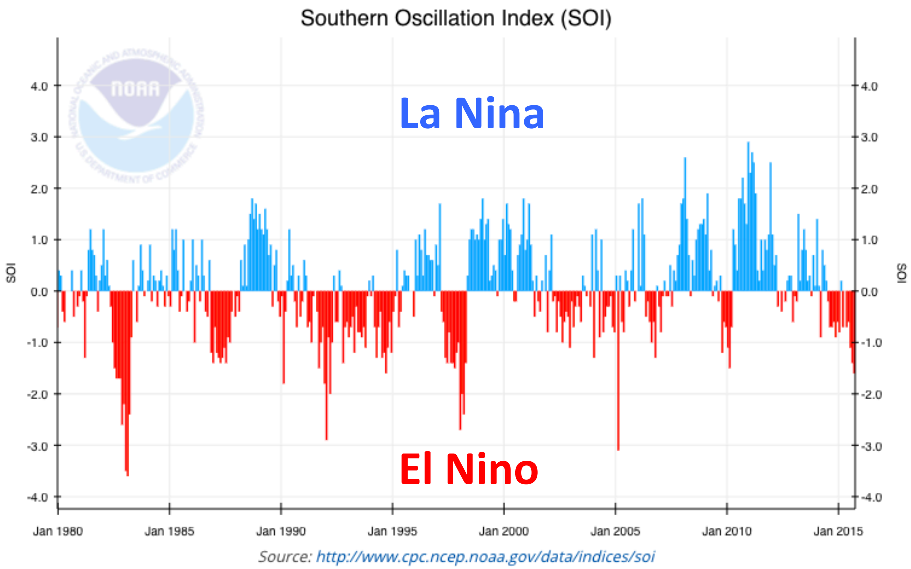
See Link: http://www.tropicaltidbits.com/analysis/models/?model=gfs&region=atl&pkg=z500a_sd&

- Can now directly compare low heights over midwestern US to high heights over east Pacific
- Standard deviation of heights relatively low in tropics, but relatively high in mid-latitudes

Standardized Anomalies

Additional applications of standardized anomalies

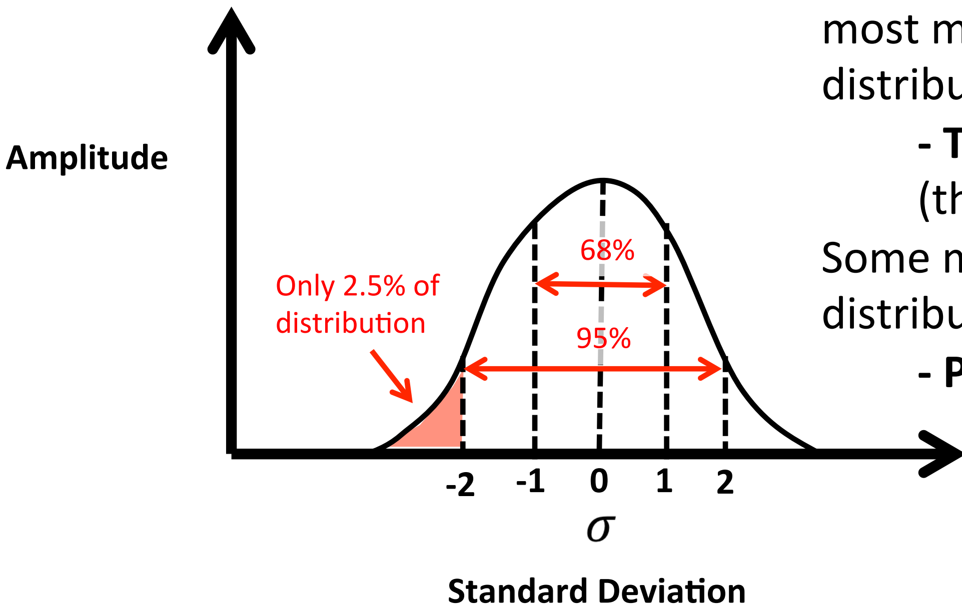
- Southern Oscillation Index (SOI)
  - ➢ Teleconnection that measures the **standardized monthly sea level pressure difference** between Tahiti and Darwin, Australia
  - ➢ Often used as a proxy for ENSO
    - Positive SOI corresponds to La Nina like conditions
    - Negative SOI corresponds to El Nino like conditions



Link:
https://www.ncdc.noaa.gov/teleconnections/enso/indicators/soi/

Normal Distribution

- Also known as a Gaussian distribution or "bell curve"



most meteorological variables have normal distributions

- Temperature, Geopotential Height

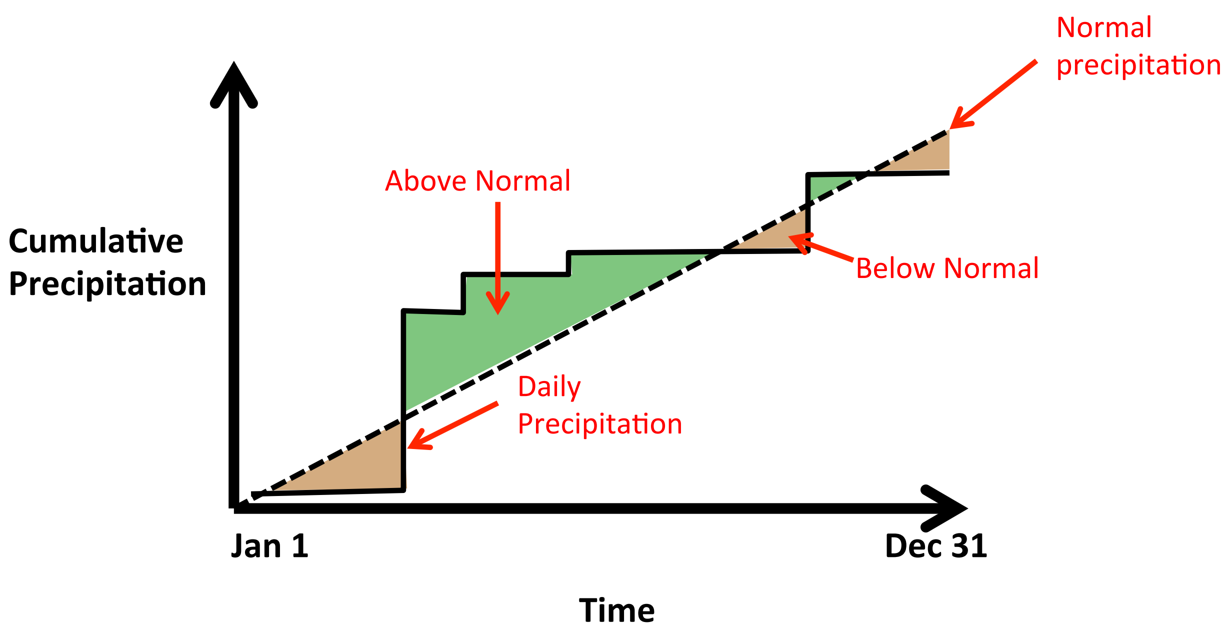
(these fields are good to "standardize")

Some meteorological variables DO NOT have normal distributions

- Precipitation, precipitable water, windspeed

Cumulative Frequency Distributions

- Depicts the cumulative total of a variable over time



URL to CFD plots from Climate Prediction Center: http://www.cpc.ncep.noaa.gov/products/global_monitoring/precipitation/global_precip_accum.shtml

Pearson (Ordinary) Correlation

Correlation measures the strength of a <u>linear</u> relationship between 2 variables

$$r_{xy} = \frac{Cov(x,y)}{\sigma_x \sigma_y} = \frac{1}{N-1}\sum_{i=1}^{N}\frac{(x_i - \bar{x})}{\sigma_x}\frac{(y_i - \bar{y})}{\sigma_y} = \frac{1}{N-1}\sum_{i=1}^{N} Z_{xi} * Z_{yi}$$

Alternative Computational Form (useful if you don't have standardized anomalies)

$$r_{xy} = \sum_{i=1}^{N}(x_i y_i) - \frac{1}{N}\left(\sum_{i=1}^{N} x_i\right)\left(\sum_{i=1}^{N} y_i\right)$$

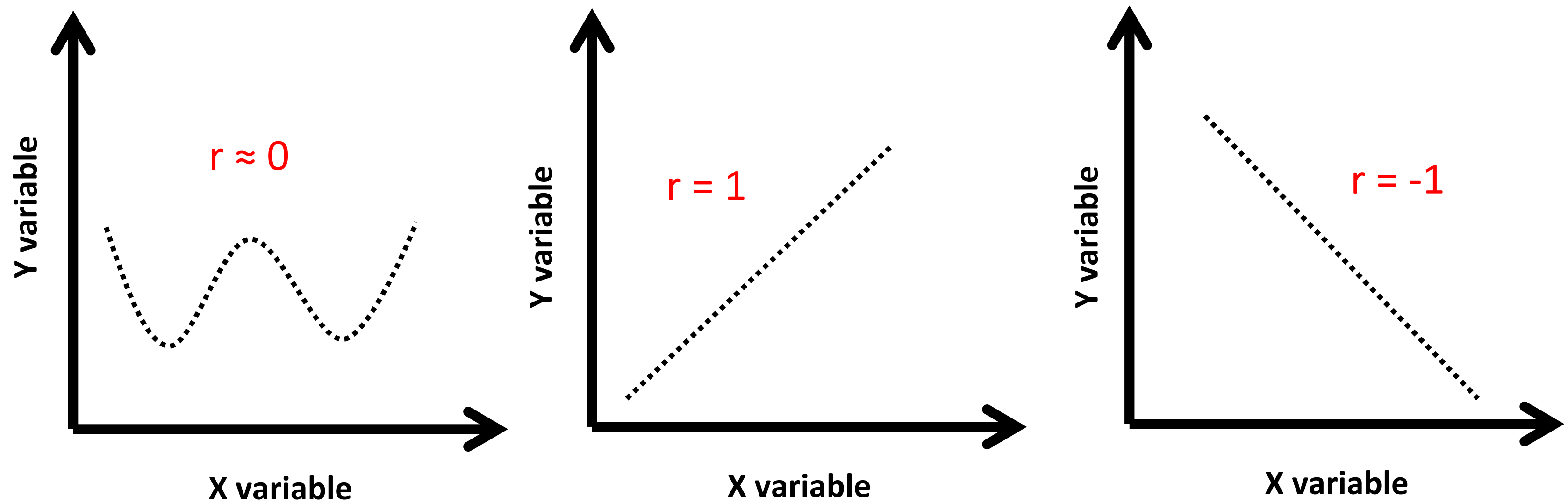
Note: $-1 \leq r_{xy} \leq 1$

-1,+1 → Indicates perfect linear association between the two variables

0 → Indicates no linear relationship between the two variables

Pearson (Ordinary) Correlation

Plot Examples



Note: even if there is no linear relationship, a strong **non-linear** relationship can exist!

IMPORTANT: Correlation can <u>suggest</u> causation, pending a physical explanation of the relationship, but it can **<u>NEVER</u>,** by itself, imply causation!

Simple Linear Regression

Let some variable x be the independent or predictor variable (i.e., SSTs)
Let some variable y be the dependent or predictand variable (i.e., # of TCs)

- A linear regression procedure chooses the line producing the least error for predictions of y given x, using a least squares approach

  → see youtube video explanation: https://www.youtube.com/watch?v=jEEJNz0RK4Q
  → Cool visualization of least squares approach http://phet.colorado.edu/sims/html/least-squares-regression/latest/least-squares-regression_en.html

- Linear regression can be used to predict linear changes in variables, and is often an important component of statistical models.
- In calculus we write a line as: $y = mx + b$
- Stasticians write line as: $\hat{y} = a + bx$

y-intercept (a)   slope (b)

$$b = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} = \frac{n\sum_{i=1}^{N}(x_i * y_i) - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{n\sum_{i=1}^{N}(x_i)^2 - \left(\sum_{i=1}^{N} x_i\right)^2} = r_{xy}\frac{S_y}{S_x}$$

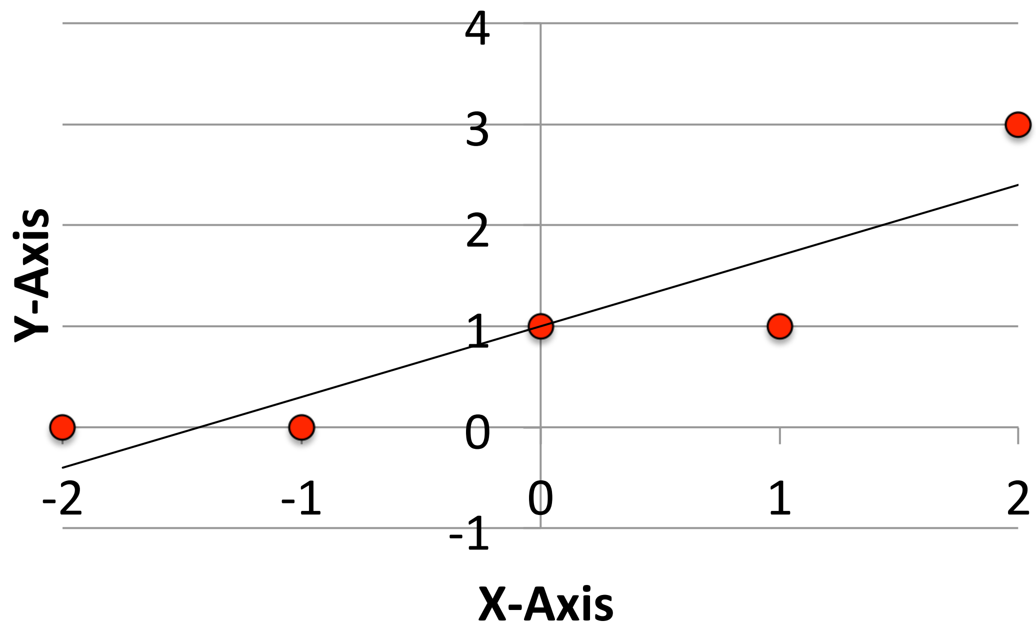
Use this form if you already know correlation coefficient

$$a = \bar{y} - b\bar{x}$$

Simple Linear Regression

Simple example

x_i	y_i
-2	0
-1	0
0	1
1	1
2	3



$x_i * y_i$	x_i^2
0	4
0	1
0	0
1	1
6	4

- Sum these values

$$\sum_{i=1}^{N} x_i = 0 \quad \sum_{i=1}^{N} y_i = 5 \quad \sum_{i=1}^{N} x_i y_i = 7 \quad \sum_{i=1}^{N} {x_i}^2 = 10$$

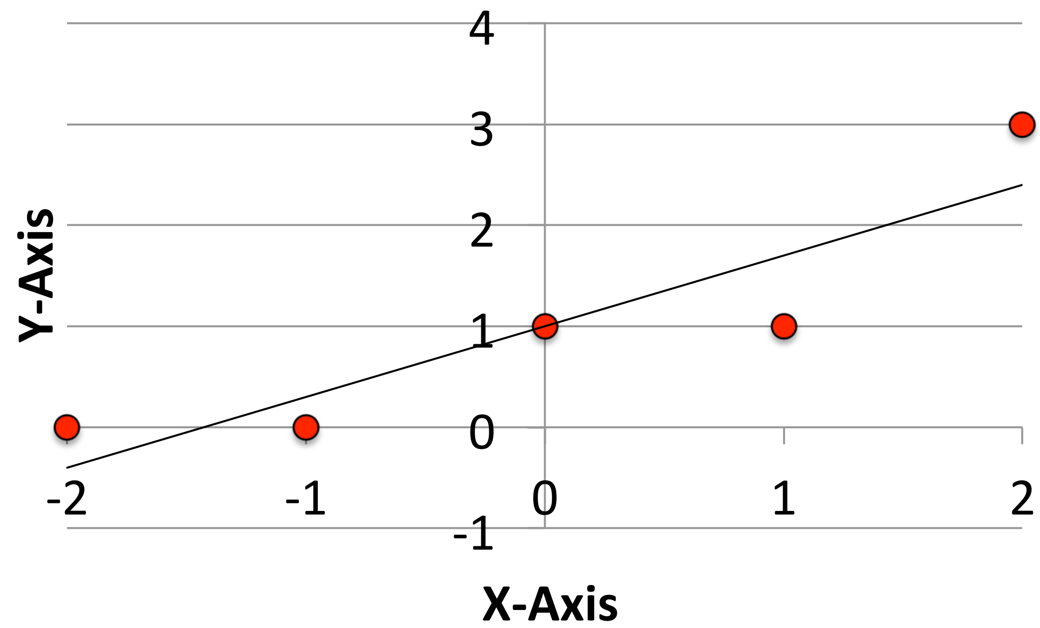
- Solve for b

$$b = \frac{n\sum_{i=1}^{N}(x_i * y_i) - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{n\sum_{i=1}^{N}(x_i)^2 - \left(\sum_{i=1}^{N} x_i\right)^2} = \frac{(5)(7) - (0)(5)}{(5)(10) - (0)^2} = 0.7$$

Simple Linear Regression

Simple example

x_i	y_i
-2	0
-1	0
0	1
1	1
2	3



$x_i * y_i$	x_i^2
0	4
0	1
0	0
1	1
6	4

- Sum these values

$$\sum_{i=1}^{N} x_i = 0 \quad \sum_{i=1}^{N} y_i = 5 \quad \sum_{i=1}^{N} x_i y_i = 7 \quad \sum_{i=1}^{N} {x_i}^2 = 10$$

- Solve for a

$$\hat{y} = a + bx$$

$$a = \bar{y} - b\bar{x} \;=\; 1 - 0.7(0) = 1$$

$$\boxed{\hat{y} = 1 + 0.7x}$$

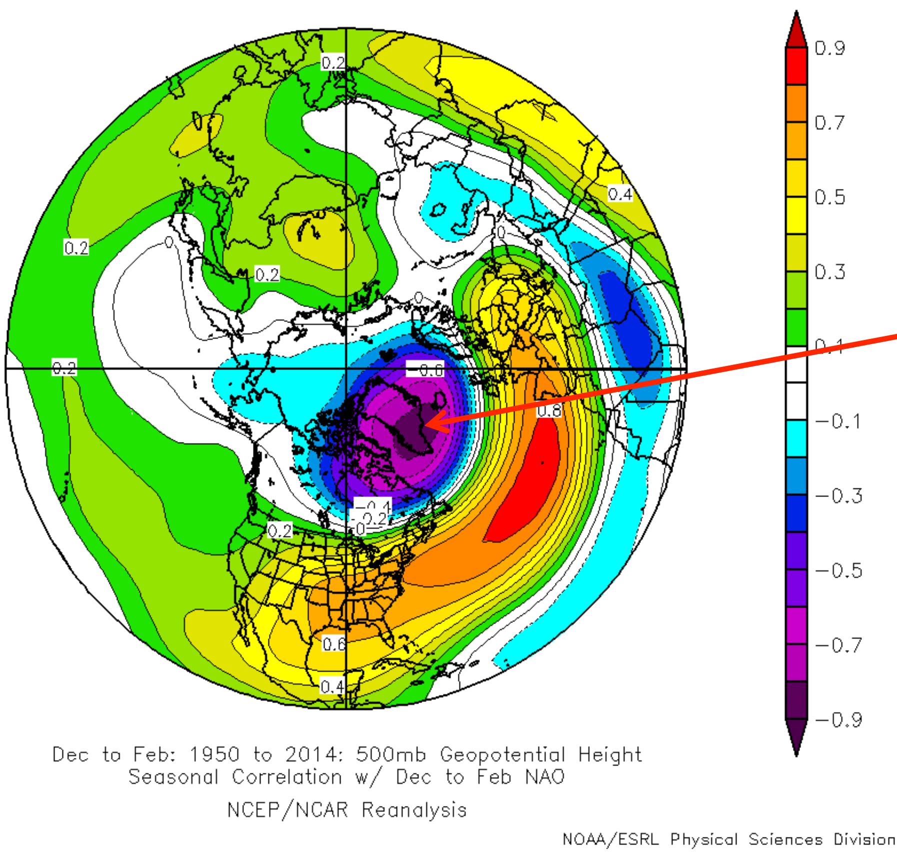
Correlation and Regression Maps

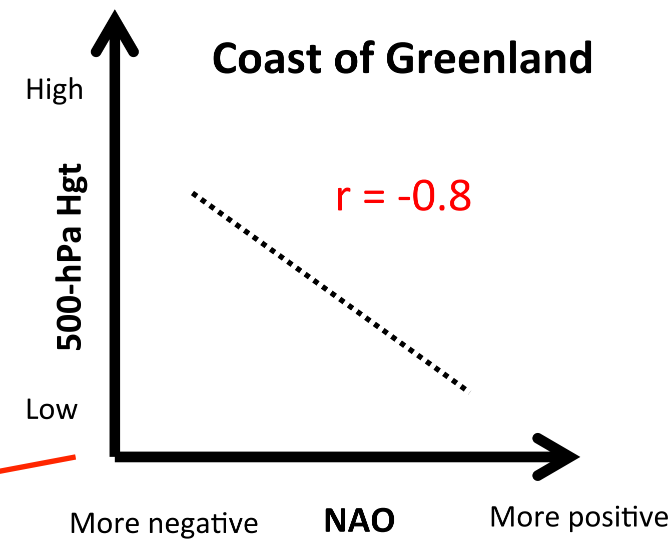
Correlation

Available on NOAA ESRL Map Room: http://www.esrl.noaa.gov/psd/data/correlation/

Example: Correlation of NAO with 500-hPa geopotential heights

At each grid point in Northern Hemisphere





Strong negative correlation

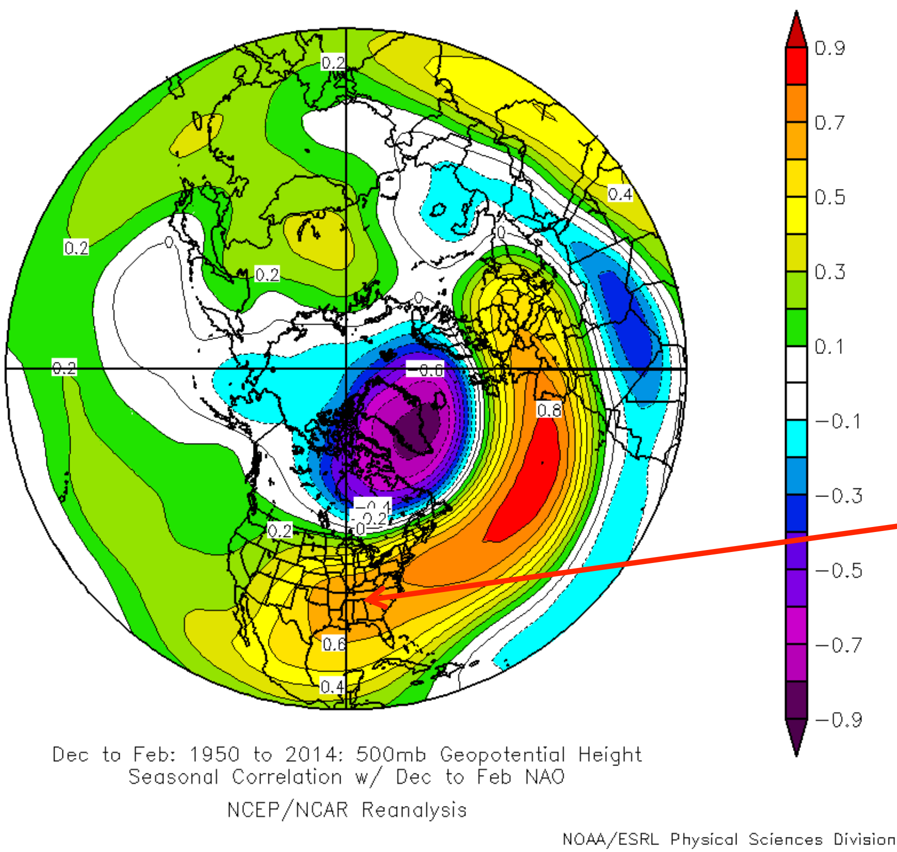
Correlation and Regression Maps

Correlation

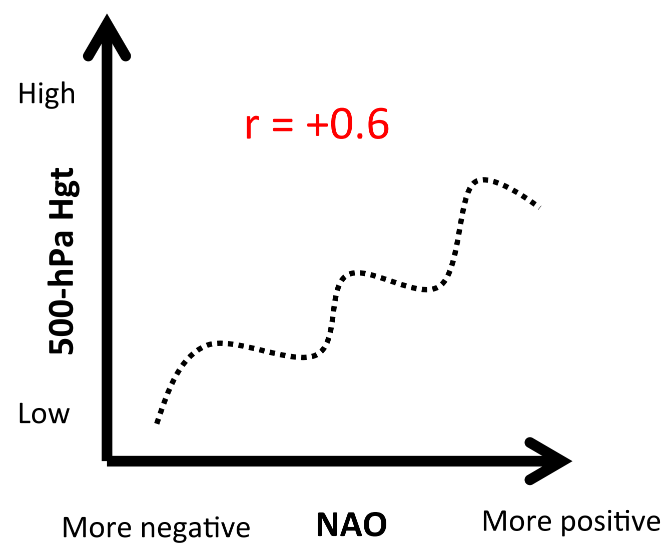
Available on NOAA ESRL Map Room: http://www.esrl.noaa.gov/psd/data/correlation/

Example: Correlation of NAO with 500-hPa geopotential heights

At each grid point in Northern Hemisphere



Eastern US



Moderate positive correlation

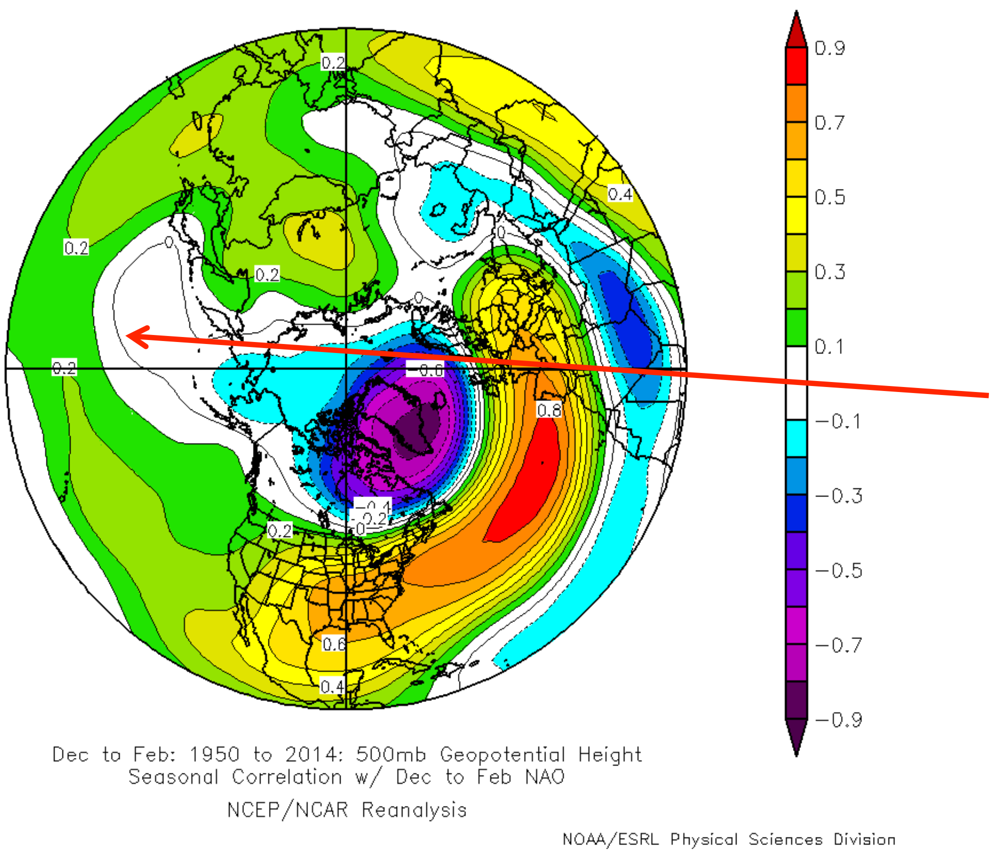
Correlation and Regression Maps

Correlation

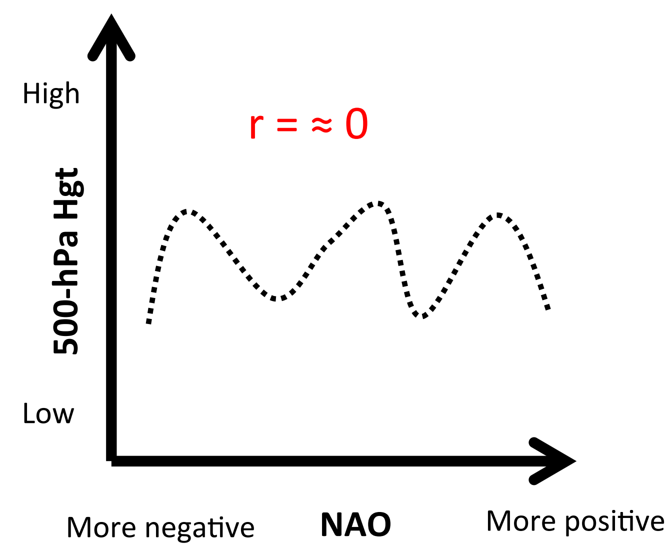
Available on NOAA ESRL Map Room: http://www.esrl.noaa.gov/psd/data/correlation/

Example: Correlation of NAO with 500-hPa geopotential heights

At each grid point in Northern Hemisphere



Central Pacific



No Correlation

Correlation and Regression Maps

Regression

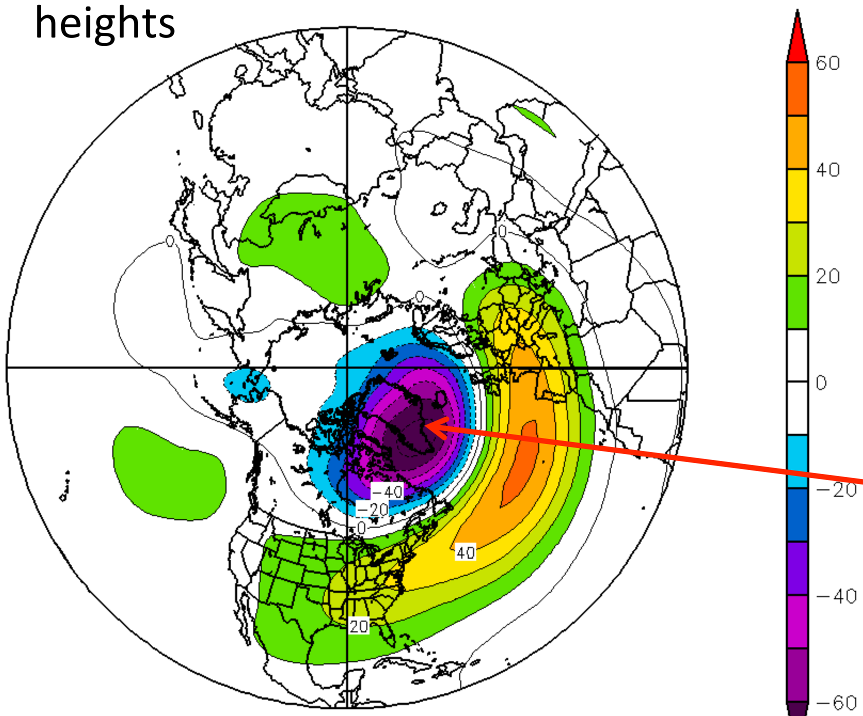
Available on NOAA ESRL Map Room: http://www.esrl.noaa.gov/psd/data/correlation/

Variable and Date Options:

○ **Correlation** ◉ **Regression**

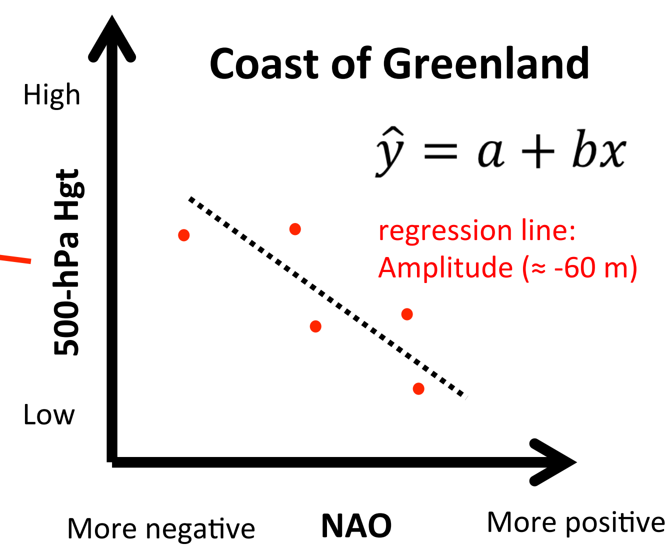
Example: Regression of NAO with 500-hPa geopotential heights

At each grid point in Northern Hemisphere



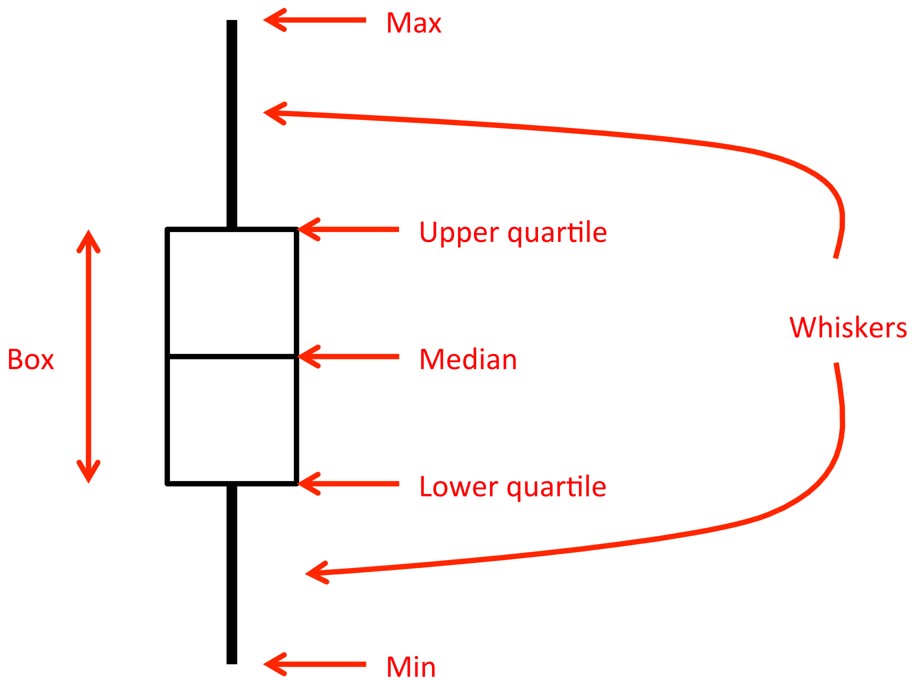
Dec to Feb: 1950 to 2014: 500mb Geopotential Height
Seasonal Regression on Geopotential Height w/ Dec to Feb NAO
NCEP/NCAR Reanalysis
NOAA/ESRL Physical Sciences Division

Note that Regression and Correlations maps show the same features, but amplitude of regression much greater than correlation which must be between -1 to 1



Simplified Statistical Summary Plots

Box and Whisker Display



Simplified Statistical Summary Plots

Histogram