Chapter 1 Basic notion in classical data analysis

The goal of data analysis: discover relations and features in a dataset.

1.1 Expectation and mean

1.1.1 Continuous variables

For a random variables x which takes on continuous values over a domain Ω, the expectation is given by an integral,

$$E(x) = \int_{\Omega} xp(x)dx \qquad (1.1)$$

where $p(x)$ is the probability density function. For any function $f(x)$, the expectation is

$$E(x) = \int_{\Omega} f(x)p(x)dx \qquad (1.2)$$

Example: For a normal distribution: $x \sim N(0,1)$, the expectation of x

$$E(x) = \int_{-\infty}^{\infty} xp(x)dx = \int_{-\infty}^{\infty} x\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx = \int_{0}^{\infty} x\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx + \int_{-\infty}^{0} x\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx = 0$$

1.1.2 Discrete variables

Let x be random variable which takes on discrete. For example, x can be the outcome of a die cast, where the possible values are $x_i = i$, with i=1,2,…6. The expectation or expected value of x_i from a population is given by

$$E[x] = \sum_i x_i p_i \qquad (1.3)$$

Where p_i is the probability of x_i occurring. If the die is fair, p_i is 1/6 for all *i,* So $E[x] = \sum_i x_i p_i$ =(1+2+3+4+5+6)/6=3.5. We also write

$$E[x] = \mu_x$$

with μ_x denoting the mean of x for the population.

Similarly, for any discrete function f(x), its expectation is

$$E[f(x)] = \sum_i f(x_i) p_i \qquad (1.4)$$

The expectation of a sum of random variables satisfies

$$E[ax + by + c] = aE(x) + bE(y) + c, \qquad (1.5)$$

where x and y are random variables, and $a, b, c,$ are constants.

In practice, one can only sample N measurements of $x(x_1, \ldots, x_N)$ from the population. The sample mean $\bar{x}$ or $<x>$ is calculated as

$$\bar{x} \equiv \langle x \rangle = \frac{1}{N} \sum_{i=1}^{N} x_i \qquad (1.6)$$

which is in general different from the population mean μ_x. As the same size increases, the same mean approaches the population mean.

The function of "mean" in MATLAB is mean(x)

1.2 Variance and covariance

Fluctuations about the mean value is commonly characterized by the variance of the population,

$$Var(x) \equiv E[(x - \mu_x)^2] = E(x^2 - 2x\mu_x + \mu_x{}^2] = E[x^2] - \mu_x{}^2 \qquad (1.7)$$

where (1.5) has been invoked. The standard deviation s is the positive square root of the population variance, i.e.,

$$s^2 = Var(x) \qquad (1.8)$$

The sample standard deviation $\sigma^2 = \frac{1}{N-1} \sum_{i-1}^{N} (x_i - \bar{x})^2$ (1.9)

As the sample size increases, the sample variance approaches the population variance. For large *N*, distinction is often not made between having N-1 or N in the denominator of (1.9).

Normalization: The goal is to make different variables to be measured and compared in the same scale, i.e., an effective way to remove the influence of units.

For example, we would like to compare two very different variables, e.g., sea surface temperature and fish population. Simply, one can't even draw their variations in a plot due to different units. So, one usually standardizes the variables before making the comparison. The standardized variable

$$x_s = (x - \bar{x})/\sigma \qquad (1.10)$$

is obtained from the original variables by subtracting the sample mean and dividing by the sample standard deviation. The standardized variable is also called the normalized variable or the standardized anomaly (where anomaly means the deviation from the mean value).

For two random variables x and y, with mean μ_x and μ_y respectively, their covariance is given by

$$Cov(x, y) = E[(x - \mu_x)(y - \mu_y)] \qquad (1.11)$$

The variance is simply a special case of the covariance, with

$$Var(x) = Cov(x, x). \qquad (1.12)$$

The sample covariance is computed as

$$Cov(x, y) = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}). \qquad (1.13)$$

The function of variance, standard deviation and covariance in MATLAB is var(x) , std(x), cov(x,y).

1.3 Skewness:

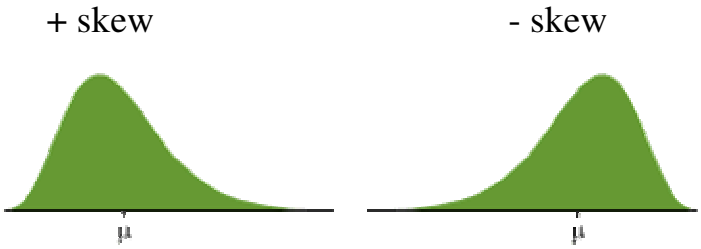
Skewness is a measure of symmetry or more precisely, the lack of symmetry.

$$skew(x) = \frac{E[(x - \mu_x)^3]}{\sigma_x^{\,3}} \qquad \textbf{(1.14)}$$

where μ_x and σ_x are the mean and standard deviation of x. As one might expect, the formula takes on a positive value if x is positively skewed and a negative value if x is negatively skewed

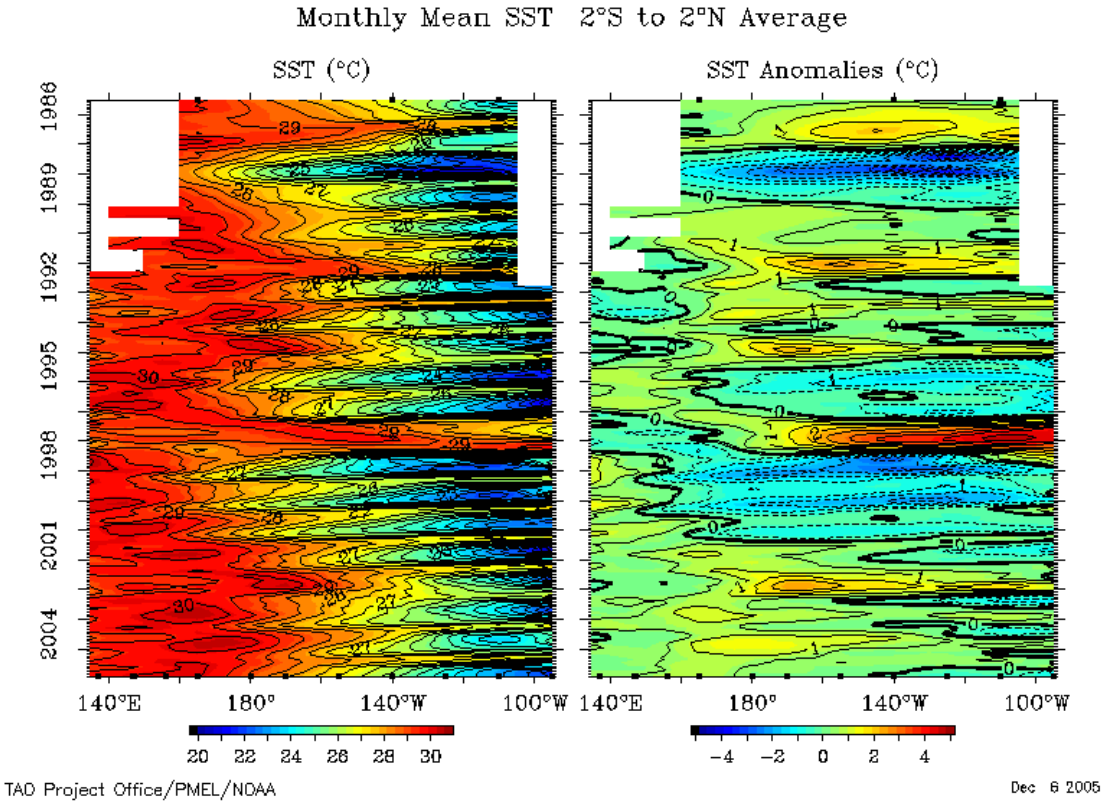
A distribution, or data set, is symmetric if it looks the same to the left and right of the center point, i.e., skew = 0. If the left tail (tail at small end of the the distribution) is more pronounced that the right tail (tail at the large end of the distribution), the skew is negative . If the reverse is true, the skew is positive. Distributions with positive skew are sometimes called "skewed to the right" whereas distributions with negative skew are called "skewed to the left."

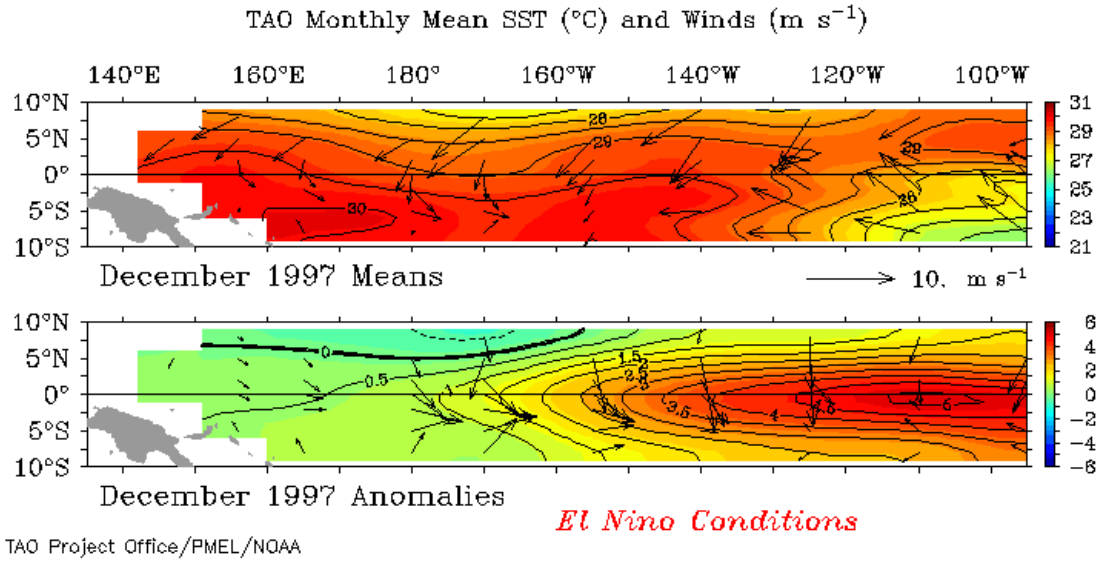
Positive vs. Negative Skewness

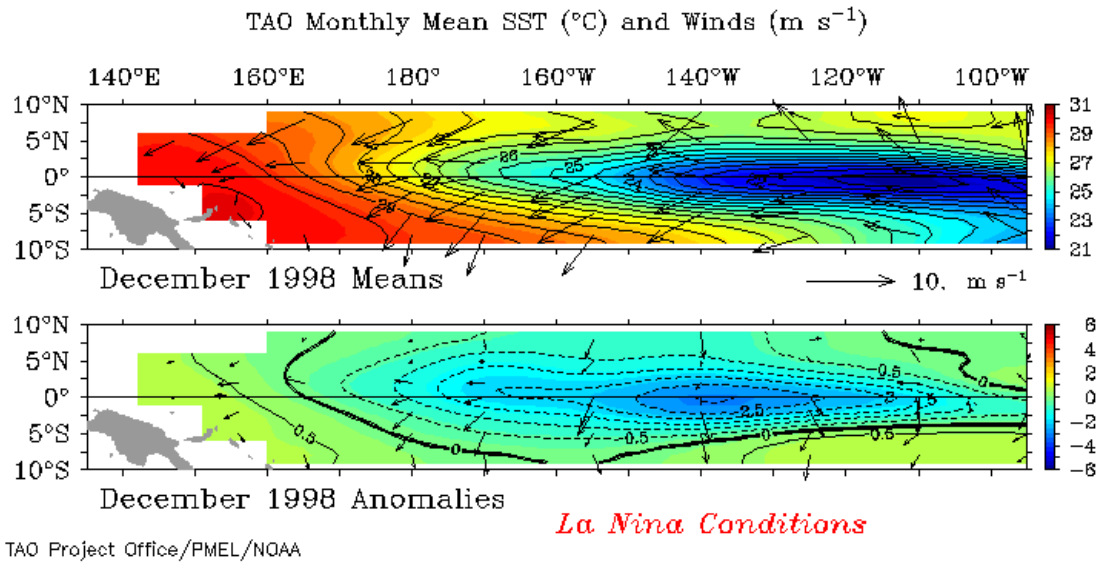


These graphs illustrate the notion of skewness. Both PDFs have the same expectation and variance. The one on the left is positively skewed. The one on the right is negatively skewed.

1.4 Examples of using these concepts to analysis practical problems.

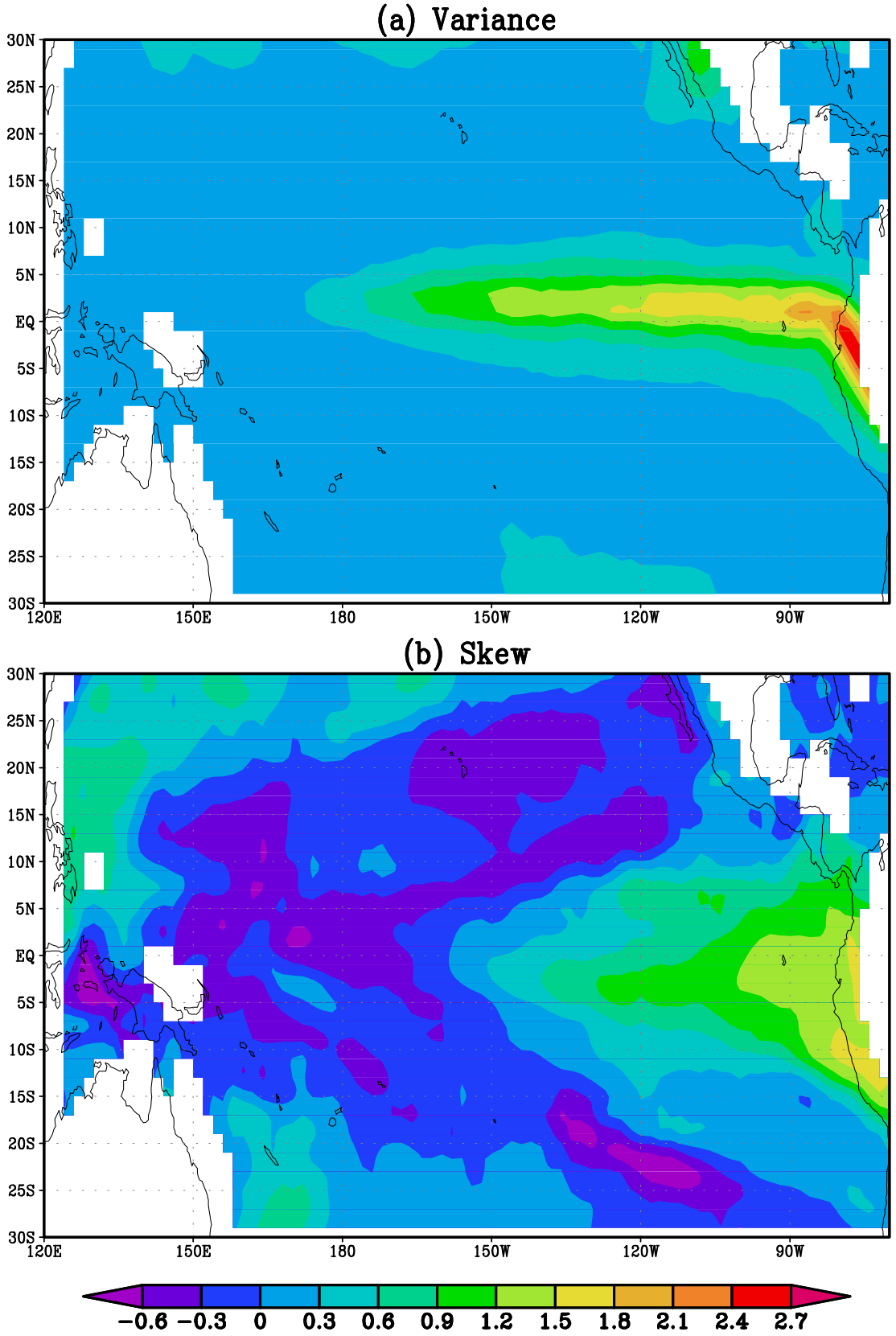

Monthly Mean SST 2°S to 2°N Average
SST (°C)
SST Anomalies (°C)
1986
1989
1992
1995
1998
2001
2004
140°E
180°
140°W
100°W
20 22 24 26 28 30
−4 −2 0 2 4
TAO Project Office/PMEL/NOAA
Dec 6 2005


TAO Monthly Mean SST (°C) and Winds (m s−1)
140°E
160°E
180°
160°W
140°W
120°W
100°W
10°N
5°N
0°
5°S
10°S
December 1997 Means
10. m s−1
December 1997 Anomalies
El Nino Conditions
TAO Project Office/PMEL/NOAA



Shown above are two cases: El Nino and La Nina. Comparing the two cases reveals: (1) the largest variations occur at the eastern Pacific ocean; (2) El Nino and La Nina are asymmetric.

These two features could be explained by the below figure. As can be seen, the largest magnitudes of variances appear over the eastern Pacific, coinciding with the above figures. Similarly, the large positive skewness occupies the equatorial eastern boundary, and small negative skewness appears in the west, indicating the asymmetry shown between El Nino and La Nina.


(a) Variance
30N
25N
20N
15N
10N
5N
EQ
5S
10S
15S
20S
25S
30S
120E
150E
180
150W
120W
90W
(b) Skew
30N
25N
20N
15N
10N
5N
EQ
5S
10S
15S
20S
25S
30S
120E
150E
180
150W
120W
90W
−0.6 −0.3 0 0.3 0.6 0.9 1.2 1.5 1.8 2.1 2.4 2.7