**Boring but important disclaimers**:

- If you are not getting this from the GitHub repository or the associated Canvas page (e.g. CourseHero, Chegg etc.), you are probably getting the substandard version of these slides Don't pay money for those, because you can get the most updated version for free at

  `https://github.com/julianmak/academic-notes`

  The repository principally contains the compiled products rather than the source for size reasons.
- Associated Python code (as Jupyter notebooks mostly) will be held on the same repository. The source data however might be big, so I am going to be naughty and possibly just refer you to where you might get the data if that is the case (e.g. JRA-55 data). I know I should make properly reproducible binders etc., but I didn't...
- I do not claim the compiled products and/or code are completely mistake free (e.g. I know I don't write Pythonic code). Use the material however you like, but use it at your own risk.
- As said on the repository, I have tried to honestly use content that is self made, open source or explicitly open for fair use, and citations should be there. If however you are the copyright holder and you want the material taken down, please flag up the issue accordingly and I will happily try and swap out the relevant material.

# OCES 3301 :

## basic Data Analysis in ocean sciences

Session 2: basic manipulations and statistics

# Outline

(Just overview here; for actual content see Jupyter notebooks)

- basic stats with basic example
- El Nino 3.4 SST data

  $\rightarrow$ demonstration of data

  $\rightarrow$ overview

  $\rightarrow$ some plotting + exercises

# Basic stats

Suppose I have some data samples as the following:

$$x_i = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$$

- **sample size** $N$, the number of samples
- **range**, largest minus smallest of sample
  $\rightarrow$ crude measure of spread

averages (but actually three of these):

1. **mode**, most frequent occurrence
2. **median**, rank these, and find the middle one
3. **mean**, THE average

# Basic stats

$$x_i = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$$

- **lower/upper (25/75 percent) quartile**, rank data, value at which 25/75 percent of data lie below
- **inter-quartile range**, the different between upper and lower quartile

  $\rightarrow$ measures spread

# Basic stats

Summary as a box-and-whisker plot

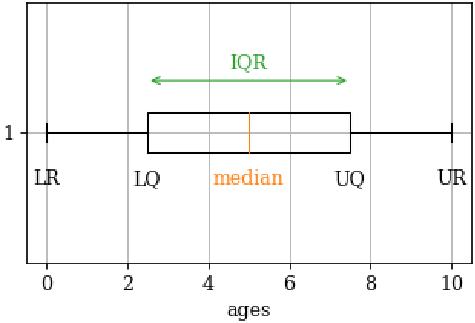



**Figure:** Nobel prize winning box plot.

# Basic stats

$$x_i = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$$

- **mean**, THE average

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$\rightarrow$ sum up, divide by number going into sum

# Basic stats

$$x_i = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$$

- (unadjusted) **variance**

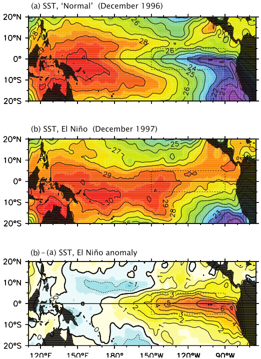
$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_N - \bar{x})^2}{N} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2,$$

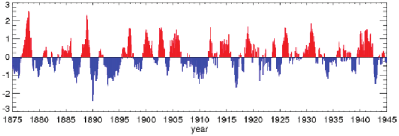
$\rightarrow$ take mean off sample, square each result, sum, divide by number going into sum

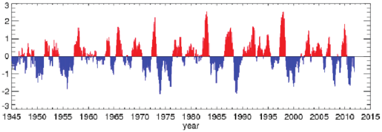
$\rightarrow$ square-root of variance is the **standard deviation (s.t.d.)**

# El-Niño (see also ENVS 3004; probably see also OCES 4001)

- “the little boy”, known to fisherman in South America for a long time
- generally starts around Christmas time
- warming in Eastern equatorial Pacific ocean
- signal in SST in modern day
  → proxy data from **corals**



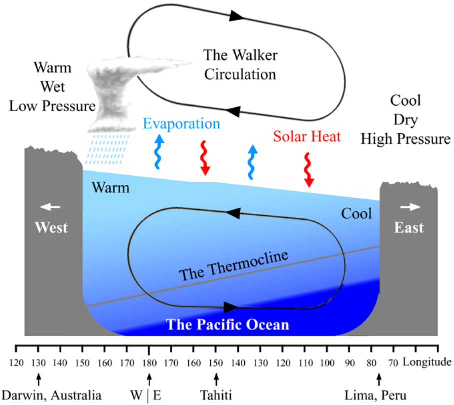
# Southern Oscillation

- discovered by Gilbert Walker (1868–1958)

  → correlation with monsoon and thus famine and drought in India

  → Companion of the Order of the Star of India in 1911
- winds change directions periodically

  → Walker circulation changes (E-W, part of N-S Hadley circulation)




(Ocean heat source moving affects atmospheric circulation)

# “Normal” + El-Niño event

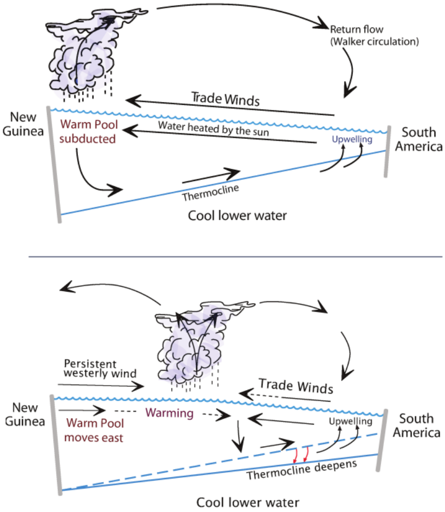



**Figure:** Schematic of ENSO, from Vallis (2019).

# El-Niño 3.4 region

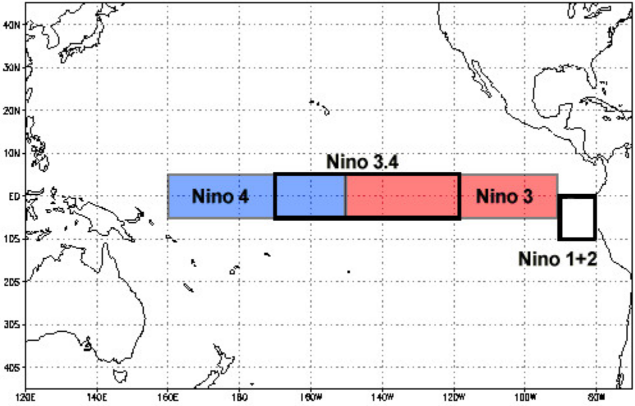



**Figure:** Pre-defined regions related to El-Niño indices. Picture probably (?) from NOAA.

# El-Niño 3.4 SST

Example of time-series data (see S07 and S08 also)

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1948 | | 2019 | | | | | | | | | |
| 1948 | -99.99 | -99.99 | -99.99 | -99.99 | -99.99 | -99.99 | -99.99 | -99.99 | -99.99 | -99.99 | -99.99 | -99.99 |
| 1949 | -99.99 | -99.99 | -99.99 | -99.99 | -99.99 | -99.99 | -99.99 | -99.99 | -99.99 | -99.99 | -99.99 | -99.99 |
| 1950 | 24.55 | 25.06 | 25.87 | 26.28 | 26.18 | 26.46 | 26.29 | 25.88 | 25.74 | 25.69 | 25.47 | 25.29 |
| 1951 | 25.24 | 25.71 | 26.90 | 27.58 | 27.92 | 27.73 | 27.60 | 27.02 | 27.23 | 27.20 | 27.25 | 26.91 |
| 1952 | 26.67 | 26.74 | 27.17 | 27.80 | 27.79 | 27.18 | 26.53 | 26.30 | 26.36 | 26.26 | 25.92 | 26.21 |
| 1953 | 26.74 | 27.00 | 27.57 | 28.04 | 28.28 | 28.12 | 27.43 | 26.94 | 27.01 | 26.87 | 26.88 | 27.00 |
| 1954 | 26.98 | 27.03 | 26.90 | 26.64 | 27.12 | 26.80 | 26.11 | 25.43 | 25.12 | 25.23 | 25.57 | 25.26 |
| 1955 | 25.61 | 25.81 | 26.22 | 26.60 | 26.66 | 26.55 | 26.15 | 25.51 | 25.28 | 24.41 | 24.25 | 24.57 |
| 1956 | 25.34 | 25.76 | 26.46 | 26.85 | 27.13 | 26.81 | 26.23 | 25.68 | 25.73 | 25.75 | 25.56 | 25.71 |
| 1957 | 26.04 | 26.54 | 27.46 | 28.23 | 28.55 | 28.36 | 28.17 | 27.69 | 27.44 | 27.42 | 27.62 | 27.90 |
| 1958 | 28.33 | 28.24 | 28.27 | 28.27 | 28.31 | 27.99 | 27.32 | 26.85 | 26.40 | 26.45 | 26.75 | 26.62 |
| 1959 | 27.07 | 27.18 | 27.47 | 27.88 | 27.70 | 27.37 | 26.44 | 26.09 | 25.92 | 26.24 | 26.04 | 26.18 |
| 1960 | 26.27 | 26.29 | 26.98 | 27.49 | 27.68 | 27.24 | 26.88 | 26.70 | 26.44 | 26.22 | 26.26 | 26.22 |
| 1961 | 26.23 | 26.56 | 26.94 | 27.36 | 27.75 | 27.67 | 26.89 | 26.19 | 25.78 | 25.71 | 26.07 | 25.97 |
| 1962 | 25.96 | 26.19 | 26.80 | 27.13 | 27.05 | 27.08 | 26.76 | 26.33 | 25.94 | 25.97 | 25.75 | 25.67 |
| 1963 | 25.77 | 26.22 | 27.18 | 27.78 | 27.63 | 27.62 | 27.78 | 27.48 | 27.40 | 27.36 | 27.47 | 27.62 |
| 1964 | 27.34 | 27.13 | 27.02 | 26.95 | 26.82 | 26.59 | 26.33 | 25.60 | 25.32 | 25.37 | 25.26 | 25.23 |
| 1965 | 25.66 | 26.19 | 26.94 | 27.38 | 27.99 | 28.09 | 27.90 | 27.97 | 28.01 | 28.17 | 28.12 | 27.96 |
| 1966 | 27.67 | 27.55 | 28.21 | 28.16 | 27.55 | 27.64 | 27.33 | 26.48 | 26.27 | 26.22 | 26.23 | 26.03 |
| 1967 | 25.88 | 26.11 | 26.50 | 26.74 | 27.35 | 27.47 | 26.97 | 26.44 | 25.86 | 25.97 | 26.08 | 25.95 |
| 1968 | 25.69 | 25.68 | 26.33 | 27.10 | 27.19 | 27.88 | 27.58 | 27.01 | 26.72 | 26.75 | 27.20 | 27.27 |
| 1969 | 27.50 | 27.86 | 27.82 | 28.13 | 28.29 | 27.69 | 27.08 | 27.02 | 27.15 | 27.34 | 27.10 | 26.98 |

**Figure:** Sample content of `elnino34_sst.data`.

# Iris data

Example of categorical + numerical data (multivariate), in `iris.csv`

- from Ronald Fisher's dataset (will see him time and again in this course)
- the "hello world" of statistics and machine learning

**Figure:** *Iris setosa, versicolor, and virginica*, in the iris dataset. Often used in machine learning (e.g. clustering analysis) and useful for demonstrating statistical concepts. Pictures from Wikipedia.

# Palmer Penguin data

Fisher was known to be a proponent of eugenics (whether he was a racist as such is ongoing debate), so if you want an alternative to iris data, there is the penguin data in `penguins.csv`

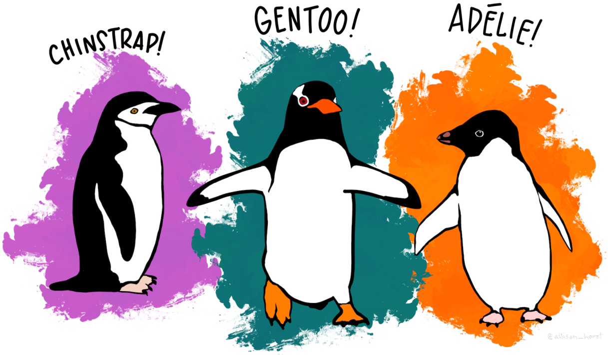



**Figure:** Artwork from Palmer Penguins dataset, by Allison Horst. See `https://allisonhorst.github.io/palmerpenguins/articles/intro.html`.

# Jupyter notebook

Go to 02 Jupyter notebook to play around with these datasets in Python