Capstone 1 - SF Bay Water Quality

Statistical Methods

For this step of my Capstone 1 project, I applied some of the inferential statistics techniques I have been learning.

Plan

"Based on your dataset, the questions that interest you, and the results of the visualization techniques that you used previously, you should choose the most relevant statistical inference techniques."

Sampling Stations

Map of sampling stations. (map generated using BatchGeo)

I divided the sampling stations into groups, based on their location, starting at station 36 (at the southern end of the South Bay). Each group, except the sixth, contains eight stations.

Station Groups

Group	Location
0	South Bay (Santa Clara County)
1	South Bay (San Mateo County)
2	Central Bay (Golden Gate)
3	San Pablo Bay
4	Suisun Bay
5	Lower Sacramento River
6	Upper Sacramento River

Ideas for Things to Examine:

1. How does mean salinity compare between station groups?

2. How does mean temperature compare between station groups?

3. Correlation between phytoplankton biovolume and

    - chlorophyll or oxygen (guess: positive as chlorophyll produces O2)
    - nutrients (guess: negative as phytoplankton consume nutrients)

4. Correlation between Sigmat-t (density) and

    - Salinity (should be positive by definition)
    - Temperature (should be negative by definition)

Difference of Means

I know that salinity is very low where the Sacramento River meets the Bay and highest at the Golden Gate where the Bay meets the ocean, but how does mean salinity compare between stations south of the Gate? They're similar; are they the same?

Also, mean temperature appears to be very similar throughout the Bay. Is it statistically the same?

Mean Salinity

Null Hypothesis: Mean salinity is essentially the same for the stations nearest the Golden Gate and to the south.

Alternative Hypothesis: Mean salinity differs between the station groups at the Golden Gate, in San Mateo County, and Santa Clara county.

I calculated mean salinity for station groups 0 - 4.

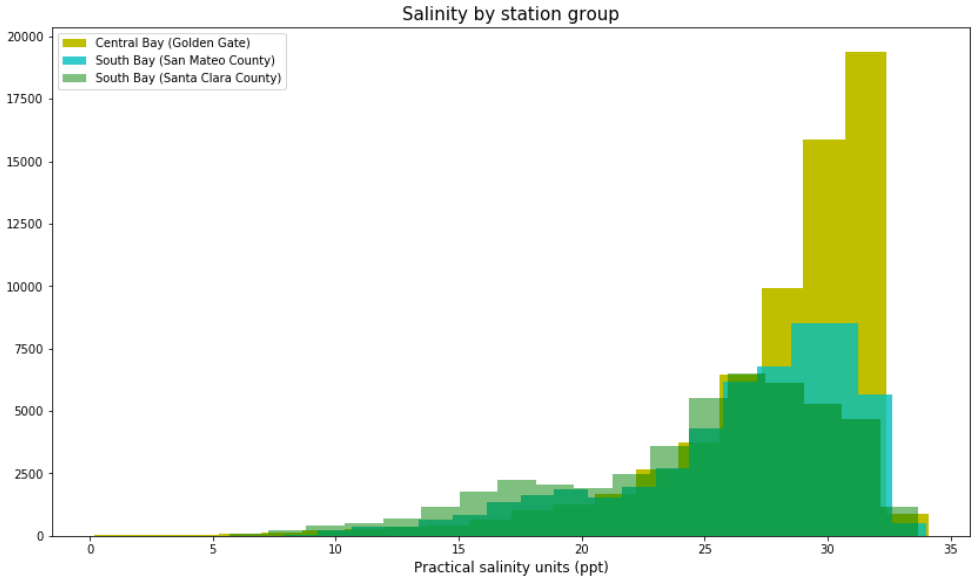
Mean Salinity

Station group	Mean salinity (practical salinity units)
0	24.78
1	26.46
2	28.06
3	17.31
4	5.78

Station groups 3 and 4 are clearly not similar. I investigated groups 0 - 2 further.

First I plotted the salinity for these station groups.

- Salinity plot by station group:



Then I checked normality and variance. The distributions are not normal (there's a tail on one side for each group) and the variances are not homogenous, but I tried a t-est anyway, using `equal_var=False`.

```
T-test result for station group 2, Central Bay (Golden Gate)
   vs group 1, South Bay (San Mateo County)
58.941584001098654
p: 0.0


T-test result for station group 1, South Bay (San Mateo County)
   vs group 0, South Bay (Santa Clara County)
50.57729469105698
p: 0.0
```

Since p < alpha (0.05) in both cases, we can say that the null hypothesis is rejected and suggest that there is a difference in the mean salinity between station groups 0, 1, and 2 - the two South Bay station groups and the Central Bay station group at the Golden Gate.

Mean Temperature

Null Hypothesis: Mean temperature is essentially the same for most station groups.

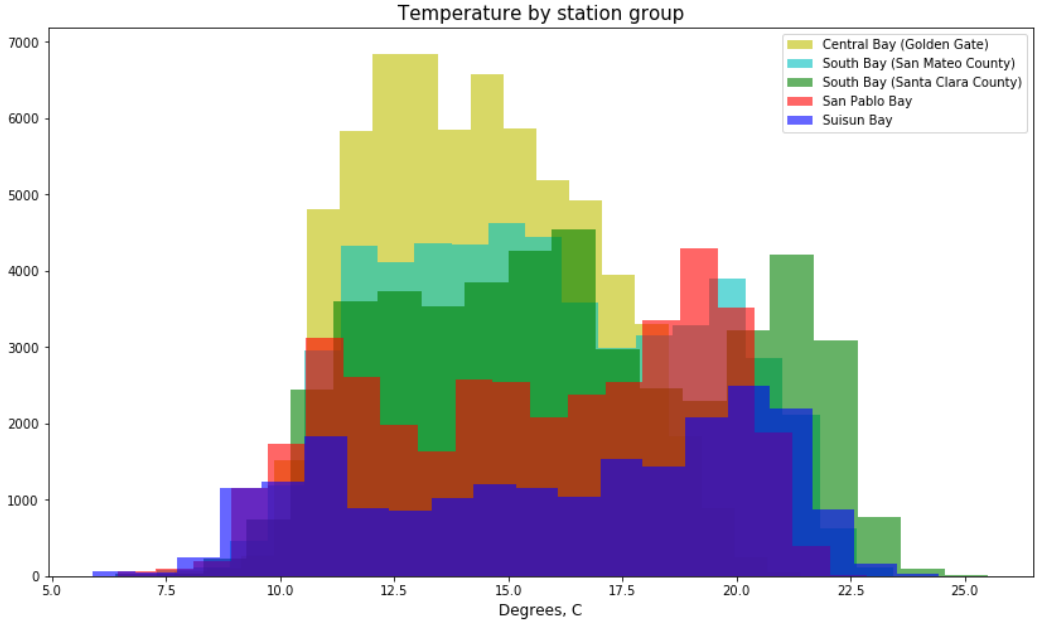
Alternative Hypothesis: Mean temperature differs between station groups.

Mean Temperature

Station group	Mean temperature (degrees C)
0 South Bay (Santa Clara County)	16.37
1 South Bay (San Mateo County)	15.67
2 Central Bay (Golden Gate)	14.36
3 San Pablo Bay	15.67
4 Suisun Bay	16.2

First I plotted the temperature for these station groups.

- Temperature plot by station group:



They overlap. The means could be the same.

Because five stations appear to have similar means, I decided to try a test that can be used for more than two sample groups.

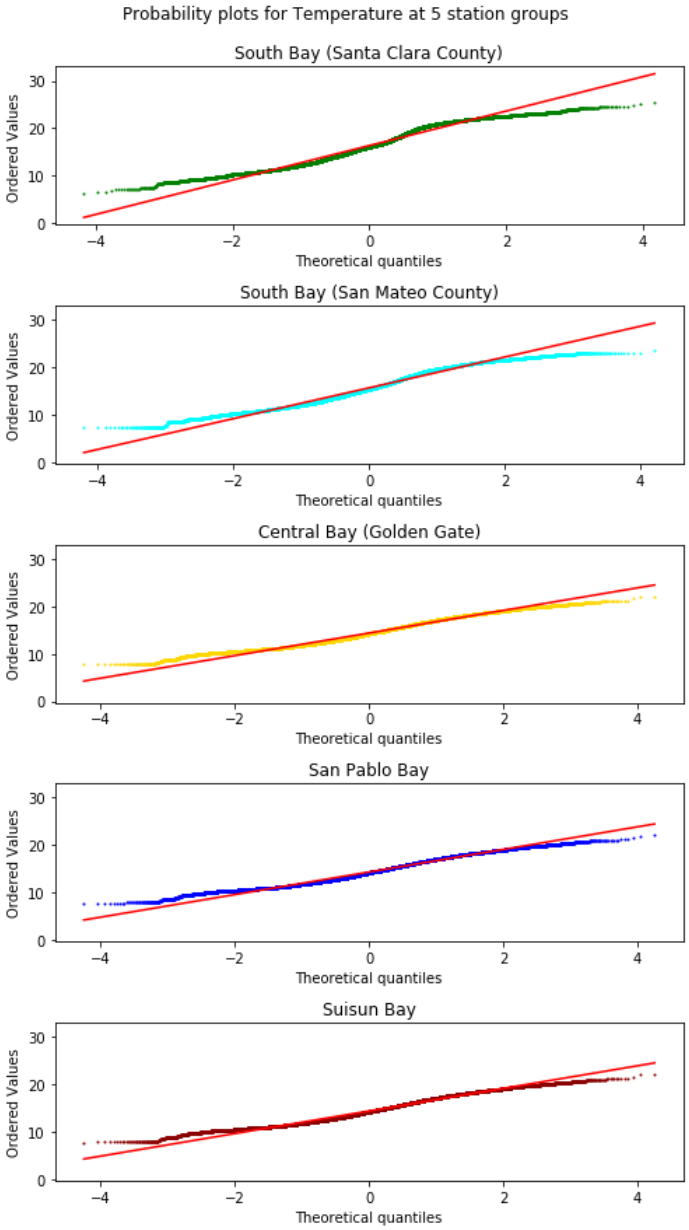
The OneWay ANOVA test can be used to compare the means from three or more groups. This test has important assumptions that must be satisfied in order for the associated p-value to be valid.

- The samples are independent.
- Each sample is from a normally distributed population.

- The population standard deviations of the groups are all equal. This property is known as homoscedasticity.

If these assumptions are not true for a given set of data, it may still be possible to use the Kruskal-Wallis H-test although with some loss of power.
The samples are independent (or as independent as they can be, all coming from the same large body of water). Next, I used probability plots to see if the samples are normally distributed. They appear to fairly close

- Probability plots:



Next, I checked the standard deviations. They're not very close.

Standard Deviation

Station group	Standard deviation for temperature
0 South Bay (Santa Clara County)	3.69
1 South Bay (San Mateo County)	3.29
2 Central Bay (Golden Gate)	2.40
3 San Pablo Bay	3.54
4 Suisun Bay	4.19

However, I did read that "if group sizes are equal, the F-statistic is robust to violation of the equal standard deviations requirement", so I checked sample size next.

Sample Size

Station group	Number ofSamples
0 South Bay (Santa Clara County)	46050
1 South Bay (San Mateo County)	53610
2 Central Bay (Golden Gate)	64894
3 San Pablo Bay	38167
4 Suisun Bay	21528

The sizes don't look close enough, so I tried the Kruskal-Wallis H-test.

For station groups 0 - 4, the p-value returned was 0.0. We cannot reject the null hypothesis; there is a significant difference in mean temperature. However, the test does not tell us where the difference is.

I ran the test again, removing station group 2 (lowest mean temperature, 14.36 degrees) from the set. This time, the p-value was *ver slightly* larger than 0.

I ran the test one more time, using only station groups 1 (South Bay (San Mateo County)) and 3 (San Pablo Bay). This time, the p-value returned was 0.0727, indicating that the mean temperature of these two sections of the bay is not statistically different.

Correlations

Next, I turned to looking for correlations between variables.

Biovolume

I tested for a correlation between:

- phytoplankton biovolume and chlorophyll
- chlorophyll and oxygen
- all nutrients and phytoplankton biovolume

I found small positive correlations between phytoplankton biovolume vs. chlorophyll (Pearsons correlation: 0.481) and biovolume vs. oxygen (Pearsons correlation: 0.333).

I was surprised to see that any correlations between biovolume and nutrients were weak, if present. The best, for biovolume vs silicate was (Pearsons) -0.308. Perhaps the quantity of dissolved nutrients in the water is sufficiently large that phytoplankton consumption has only a small effect.

Depth

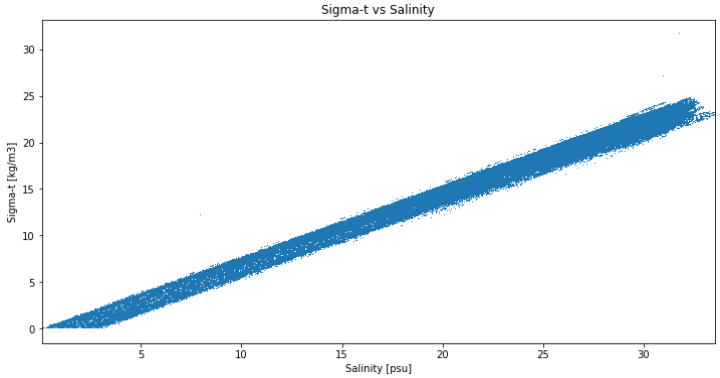
I tested for any correlation between depth and temperature or depth and suspended particulate matter (SPM). I found no correlation.

Sigma-t

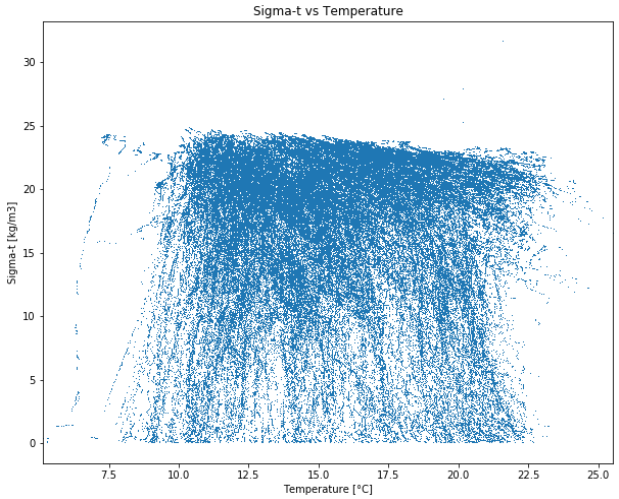
I then tested for a correlation between Sigma-t and salinity or temperature. Sigma-t is defined as "a measure of the density of the water, which is calculated as a function of salinity and temperature. Density increases with increasing salinity and decreasing temperature."

There should be a positive correlation between Sigma-t and salinity. There should be a negative correlation between Sigma-t and temperature. I found a very clear correlation between Sigma-t and salinity, but I did not find any correlation to temperature. Perhaps the San Francisco Bay never gets cold enough to show this.

- Salinity vs Sigma-t:



- Temperature vs Sigma-t:



In Summary

In summary, I can conclude that:

Mean salinity varies significantly between sections (groups of stations) across the Bay.

Mean salinity varies significantly between sections (groups of stations) across the Bay. However, the mean temperatures of San Pablo Bay and the upper end of the Souh Bay (in San Mateo County) are significantly similar.

Sigma-t is strongly correlated to salinity. Other than this, I found very few significant correlations between features.