Session 6: R check-in part 2, calculating quantiles

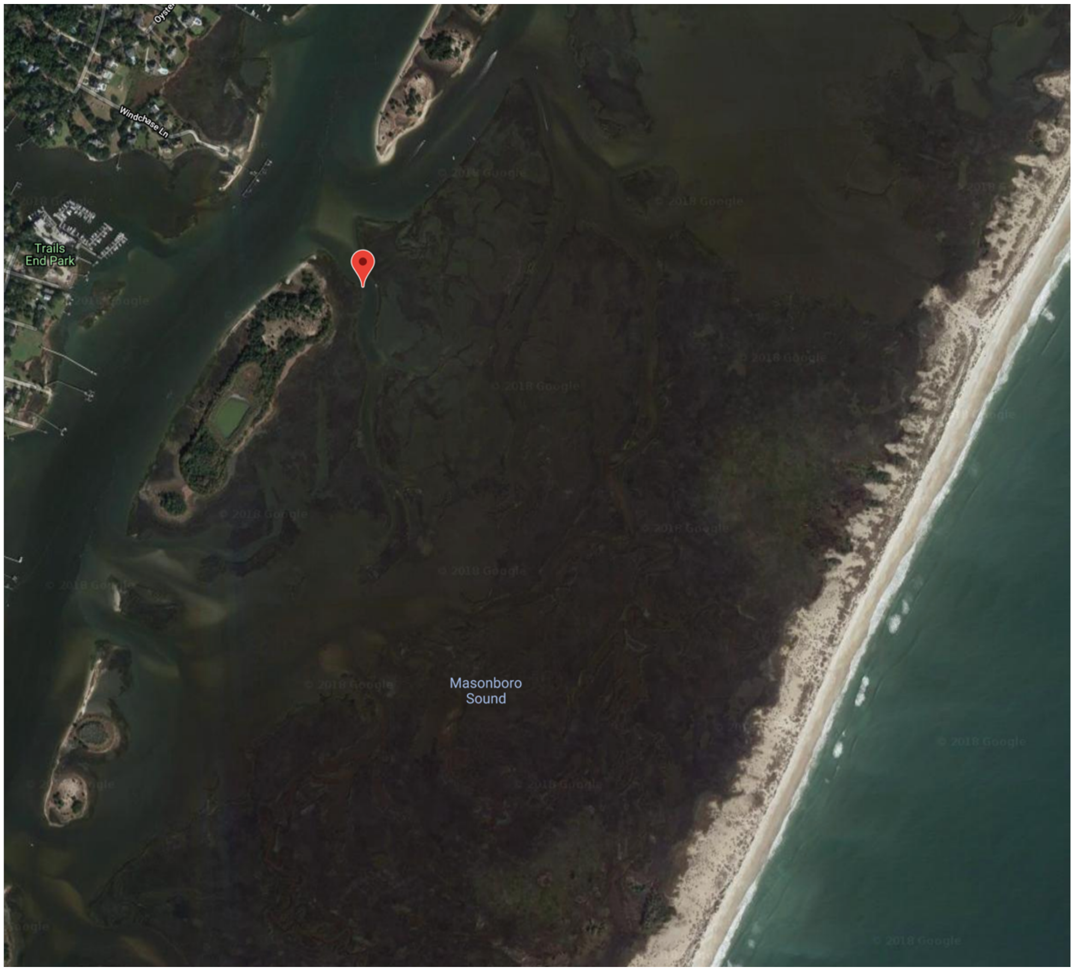
Natalie Nelson, PhD, Biological & Agricultural Engineering, NCSU

02/23/2018

1 Data

Data analyzed in this session are from the NOAA National Estuarine Research Reserve's System Wide Monitoring Program and include water depth in meters (Depth_m), dissolved oxygen in mg/L (DO_mgl), salinity in practical salinity units (Sal_psu), and water temperature in degrees Celsius (Temp_C). These measurements were taken approximately every 15 minutes from February 2017 - February 2018 in Masonboro Sound on the NC coast. The data can be accessed from "System Wide Monitoring Program's webpage".

The data are saved in one .csv file named "NOCRCWQ.csv"

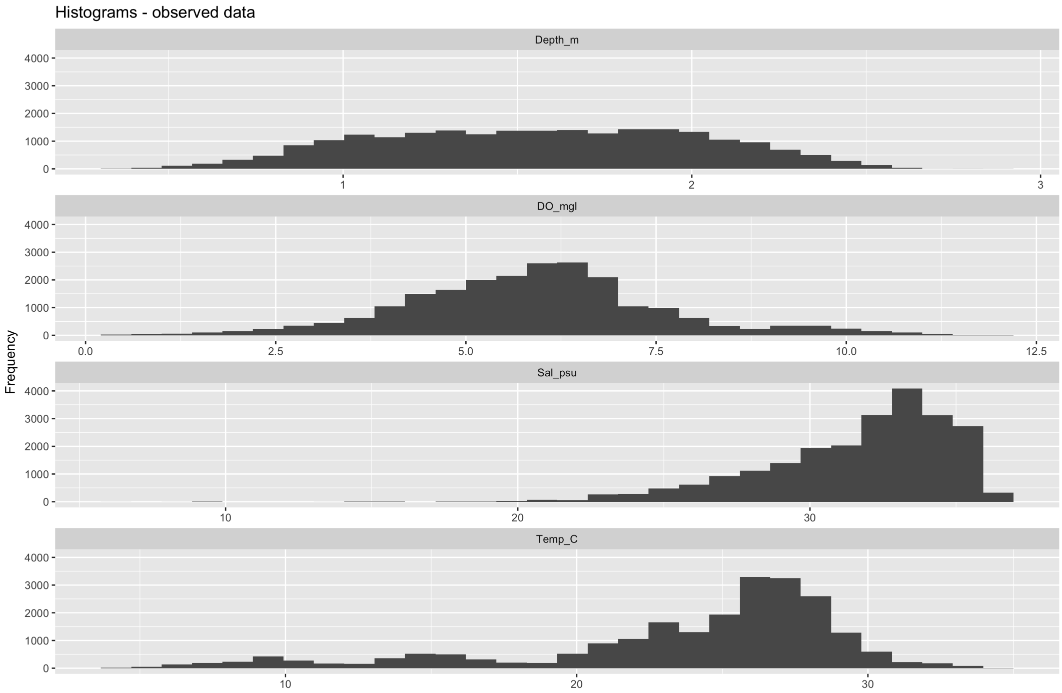


Map of the data collection site (red pin) in a tidal creek of the Masonboro Sound, NC.

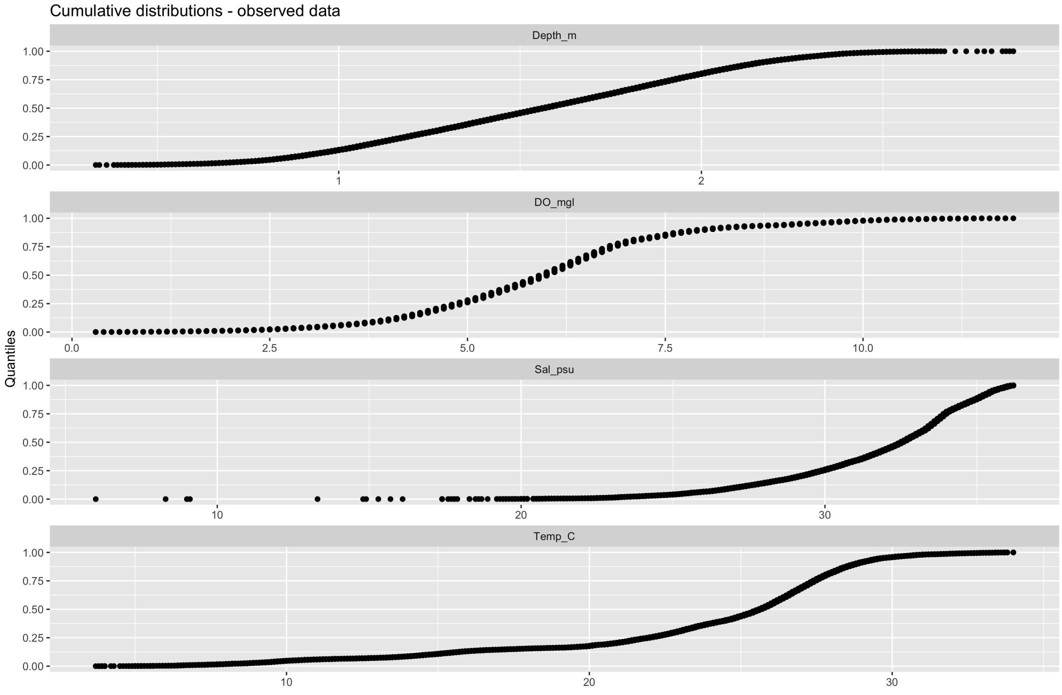
2 Final outputs

Plot 1 - histograms of observed data

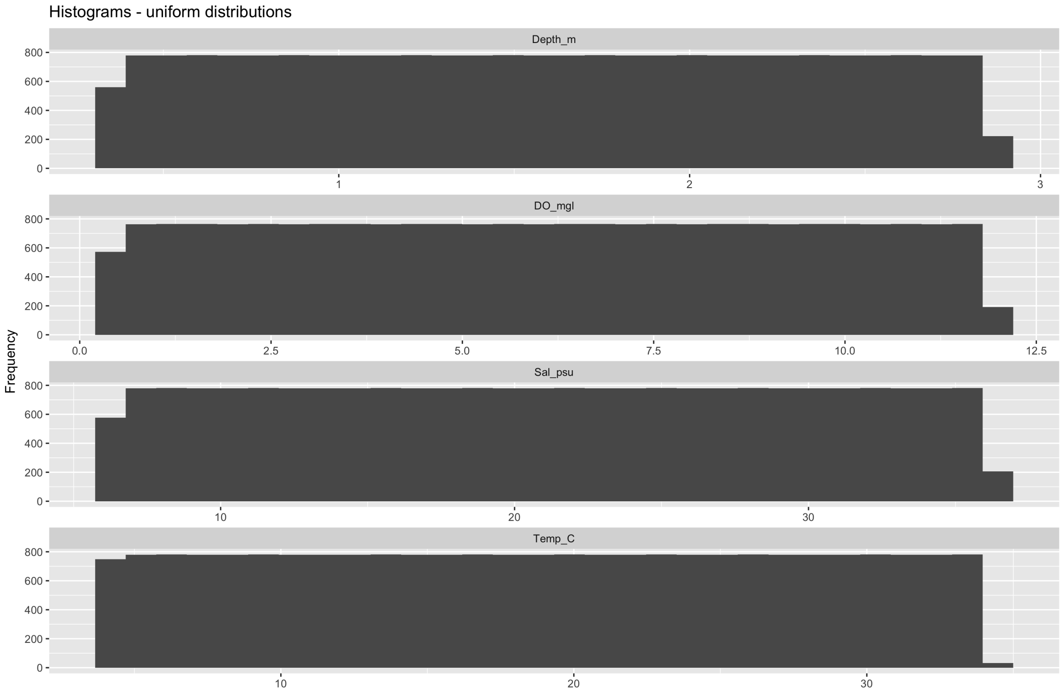
```
ggplot(d, aes(x = Measurement))+
  geom_histogram()+
  facet_wrap(~Variable, scales = "free_x", ncol = 1)+
  xlab("")+
  ylab("Frequency")+
  ggtitle("Histograms - observed data")
```



Plot 2 - cumulative distribution functions of observed data

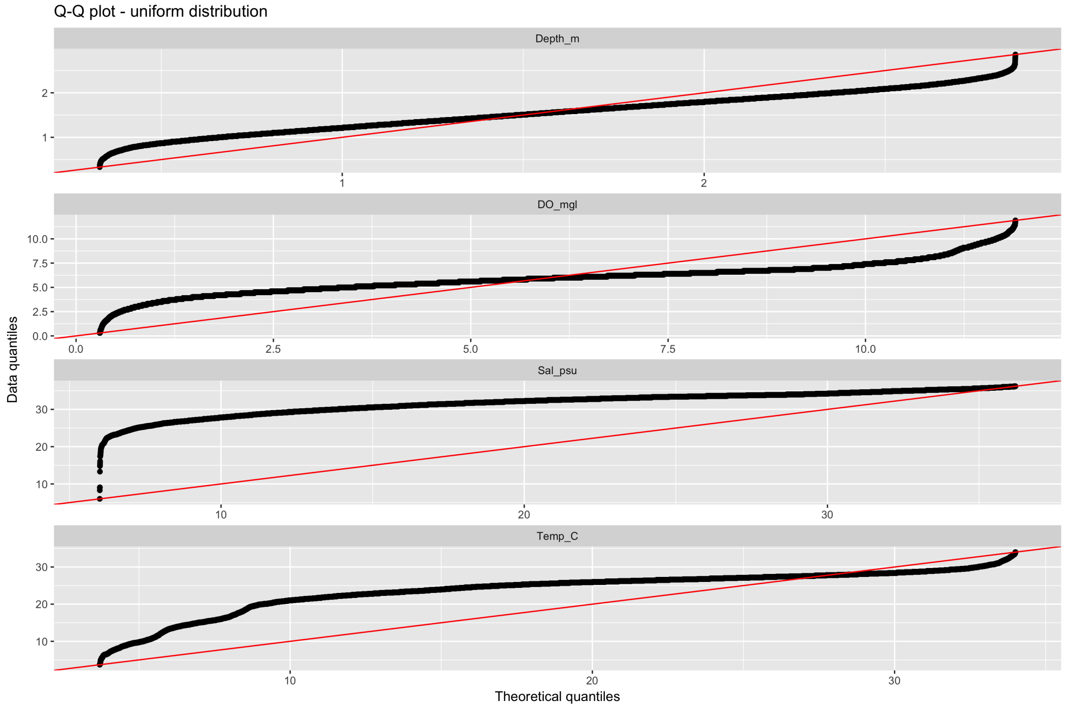


Plot 3 - histograms of values created by fitting a uniform distribution to our data



Plot 4 - Q-Q plot of observed and uniform quantiles

```
ggplot(d, aes(x = u, y = Measurement))+
  geom_point()+
  geom_abline(slope = 1, intercept = 0, color = "red")+
  facet_wrap(~Variable, scales = "free", ncol = 1)+
  xlab("Theoretical quantiles")+
  ylab("Data quantiles")+
  ggtitle("Q-Q plot - uniform distribution")
```



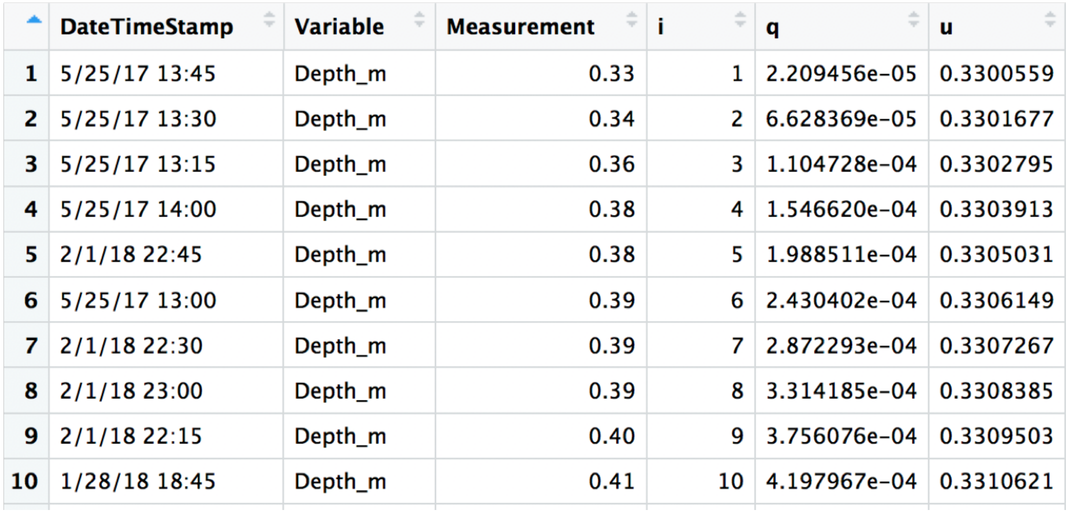
3 Workflow

1. Start by collecting your observed data into two columns (`Variable` and `Measurement`), and then remove rows with NAs.

2. Calculate the quantiles of your observed data with the formula i = q(n+1).
   2.1. Sort your data using a function described in your `dplyr` cheat sheet (you'll have to find the function by looking through the sheet). Be sure to include `.by_group = TRUE` as an argument in the function.
   Hint: if you want to test a function to see how it works, you can always create a "dummy" object to apply it to. To create a dummy object, you can use `c()`, e.g. `x <- c(1, 5, 6, 9, 10, 2)`.
   2.2. Create a new column (`i`) that contains index values. The function `row_number()` will create values from 1 to n.
   2.3. Create a new column (`q`) in which you calculate your quantiles.

3. Create a new column (`u`) in which you calculate what the quantiles of your data would be if they came from a uniform distribution (= theoretical quantiles). Remember the formula for our cumulative distribution function is: F(x) = (x-a)/(b-a), where a = minimum and b = maximum.

Before creating your plots, the head of your data should look like:



	DateTimeStamp	Variable	Measurement	i	q	u
1	5/25/17 13:45	Depth_m	0.33	1	2.209456e-05	0.3300559
2	5/25/17 13:30	Depth_m	0.34	2	6.628369e-05	0.3301677
3	5/25/17 13:15	Depth_m	0.36	3	1.104728e-04	0.3302795
4	5/25/17 14:00	Depth_m	0.38	4	1.546620e-04	0.3303913
5	2/1/18 22:45	Depth_m	0.38	5	1.988511e-04	0.3305031
6	5/25/17 13:00	Depth_m	0.39	6	2.430402e-04	0.3306149
7	2/1/18 22:30	Depth_m	0.39	7	2.872293e-04	0.3307267
8	2/1/18 23:00	Depth_m	0.39	8	3.314185e-04	0.3308385
9	2/1/18 22:15	Depth_m	0.40	9	3.756076e-04	0.3309503
10	1/28/18 18:45	Depth_m	0.41	10	4.197967e-04	0.3310621