HW#1

Bernard Romey

Due: Monday, January 19, 2015

Contents

0.1 Task 1. Numerically summarize the data and answer the questions.

Specifically,

- Summarize each numerical variable using commonly used summary statistics (e.g., mean, standard deviation, variance, etc.) (Table 1).

What is variance?

Variance is the average of the squared differences each observation is from the mean.

- Rank variance among all numerical variables (Table 1).

Sodium has the highest variance of all water quality variables (2.76x108) while pH has the lowest variance at 0.2 (Table 1). Comparing the means and medians for all variables, it looks like most of the distributions are skewed to the right, with the exception of pH. The largest difference between maximum and minimum observations is the total nitrogen concentration at 89736.2 mg/L.

Can you use variance to determine which variable varies most? If not, why?

No, because it it not standardized. Each variable may have a different measureing unit, therefore you would first need to convert it to a Z-score, then you can compare variances.

- Characterize COND, PH, NTL, TURB, AG_TOT using summary statistics by "ECO3".

Are there any ecoregion-specific patterns?

Yes, it looks like the water quality mean peramaters for the MT site are lower than both PL and XE (Table 2), with the PL site having the highest mean readings for all five variables.

Turbidity, total nitrogen, and conductivity are much different for the MT ecoregion when compared to both the PL and XE ecoregions. This may have something to do with the percent of agriculture in each ecoregion.

How many sites have % forest cover in watersheds greater than 80% in each ecoregion (ECO3)?

The MT ecoregion has 238 sites, the XE has 12, and the PL has 2 with a percent forest cover in watersheds greater than 80% (Table 3). This may help explain the differences in water quality patterns for each specific ecoregion in table 2.

How does conductivity (COND) and total nitrogen concentration (NTL) in these heavily forested sites differ among the 3 ecoregions?

Mean total nitrogen concentration is higher in the XE ecoregion than both the MT and PL ecoregion at 368.3, 275.5, and 142.5 mg/L, respectively (Table 4). Mean conductivity for all three ecoregions also exhibits this same pattern.

0.2 Task 2. Graphically summarize the data. Specifically,

- Make separate plots: 1) boxplot, 2) histogram, 3) plot showing the mean and 95% confidence intervals of COND for each ecoregion (ECO3).

The three requested plots (boxplot, histogram, and error bar plot) are shows in figures 1, 2, and 3, respectively.

The box plot does a good job of summarizing the distribution of conductivity for each ecoregion (Figure 1). It also shows the outliers and indicates to what extent the distribution of observations are skewed. Since most variables do not have negative values, this would explain why the distributions are all skewed to the right. Conductivity distribution is the largest for ecoregion PL, while ecoregion MT has the smallest distribution. The IQR for Mt is the smallest of each of the three ecoregions, while the PL ecoregion has the largest IQR.

The histogram in figure 2 shows the frequency distribution of conductivity for each ecoregion. The majority of conductivity observations at all three ecoregion are primarily below 2500 uS/cm, with the highest frequency below 1000 uS/cm.

The 95% confidence intervals for conductivity show that the variance for ecoregion PL is larger than both XE and MT (Figure 3). The MT site has the lowest mean conductivity and the smallest variance of all three ecoregions.

- Make x-y scatter plots between log-transformed TSS and log-transformed TURB for each ecoregion (ECO3) with fitted lines.

Turbidity and total suspended solids were observed in each ecoregion. For comparison, both turbidity and total suspended solids observations were log transformed. The PL ecoregion had the largest rate of increase (slope 1.22), while the XE was slightly less (1.09), and the MT ecoregion had the lowest rate at 0.41 (Figure 4).

Table 1. Summary statistics (mean, median, variance, standard deviaton, maximum, minimum) for numerical water quality variables (CA: Ca++ concentration (mg/L), COND: conductivity (µS/cm), Sodium: Na+ concentration (mg/L), ANC: Acid neutralizing capacity (mg CaCO3/L) , CL: Cl- concentration (mg/L), DOC: dissolved organic carbon (mg/L), NH4: NH4+ concentration (mg/L), NO3: NO3- concentration (mg/L), NTL: Total nitrogen concentration (mg/L), SIO2: dissolved silica concentration (mg/L), TSS: Total suspended solids (mg/L), TURB: Turbidity (NTU), PH) ranked by variance.

Factor	Mean	Median	Max	Min	Var	SD
Sodium	2021.7	323.9	77991.9	9.2	27616957.5	5255.2
NTL	767.9	188.5	89750.0	13.8	14485662.0	3806.0
CA	2252.3	1186.3	25693.5	27.0	11759923.5	3429.3
ANC	2610.0	1929.8	18662.0	26.1	5694973.9	2386.4
CL	534.8	76.6	33112.7	0.0	3466850.2	1861.9
COND	523.7	226.9	9790.0	9.2	689110.1	830.1
NO3	27.1	2.9	5989.9	0.0	62027.9	249.1
TURB	16.1	0.9	6126.0	0.0	42429.1	206.0
TSS	23.6	3.3	2988.0	0.0	12885.6	113.5
NH4	2.5	0.6	393.4	0.0	332.3	18.2

Factor	Mean	Median	Max	Min	Var	SD
SIO2	19.4	16.6	92.4	0.2	171.9	13.1
DOC	3.0	1.6	28.8	0.2	12.3	3.5
PH	8.0	8.0	9.9	6.1	0.2	0.5

Table 2. Variable means for COND: conductivity (µS/cm), pH, NTL: Total nitrogen concentration (mg/L), TURB: Turbidity (NTU), AG_TOT: % of agriculture in watershed in watershed by ecoregions (ECO3).

ECO3	COND_Ave	pH_Ave	NTL_Ave	TURB_Ave	AgTOT_Ave
MT	196.6	7.9	184.4	1.8	0.7
PL	1474.2	8.2	1683.9	38.6	46.1
XE	586.1	8.1	1533.7	34.4	4.1

Table 2: Mean conductivity, pH, total nitrogen concentration, turbidity, and % of agrigulture in watershed

ECO3	Number_site
MT	238
PL	2
XE	12

Table 3: Number of % of forestland in watershed sites within each acoregion (ECO3) that have more than 80% forest cover.

ECO3	avgNTL	avgCOND
MT	142.5	143.3
PL	275.5	234.8
XE	368.3	378.6

Table 4: Average COND: conductivity (µS/cm), and NTL: Total nitrogen concentration (mg/L) for each acoregion (ECO3) that have more than 80% forest cover.

Insert plot into word that wont convert

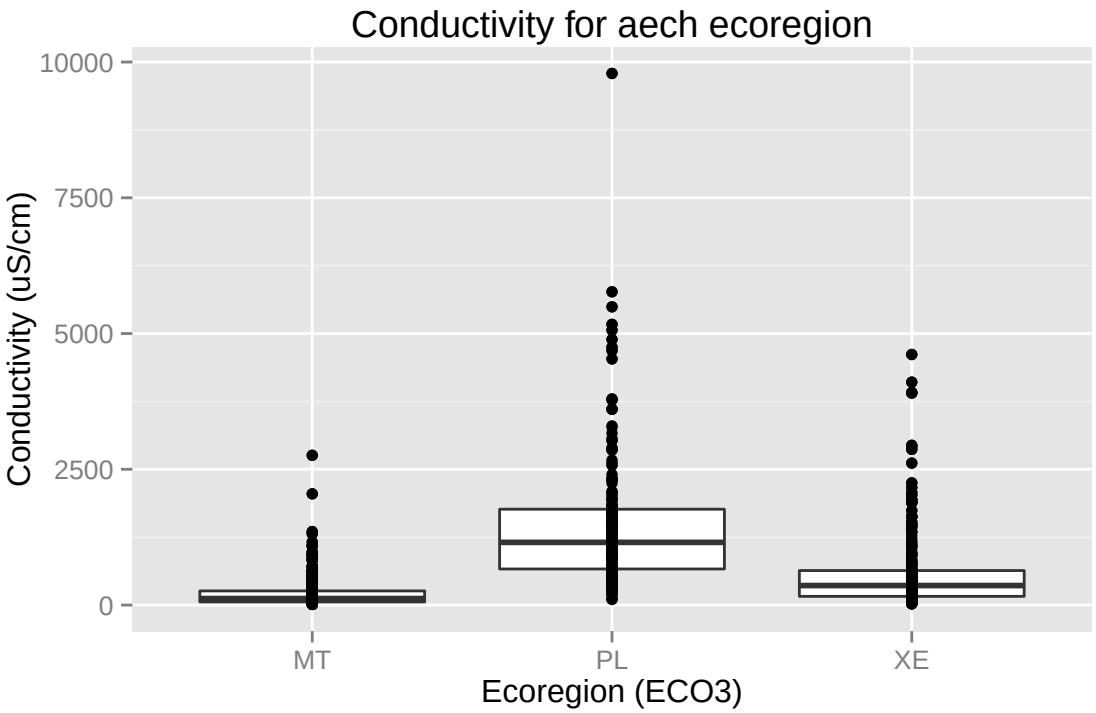

Figure 1: Box plot of conductivity for each ecoregion (ECO3)

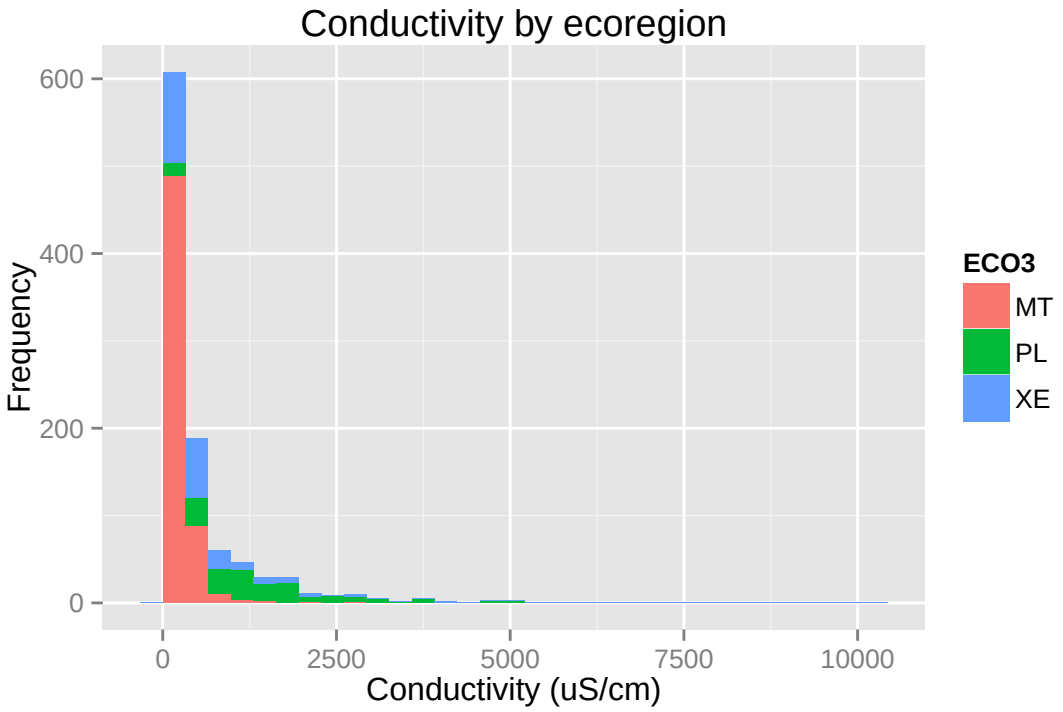

Figure 2: Conductivity histogram by ecoregion

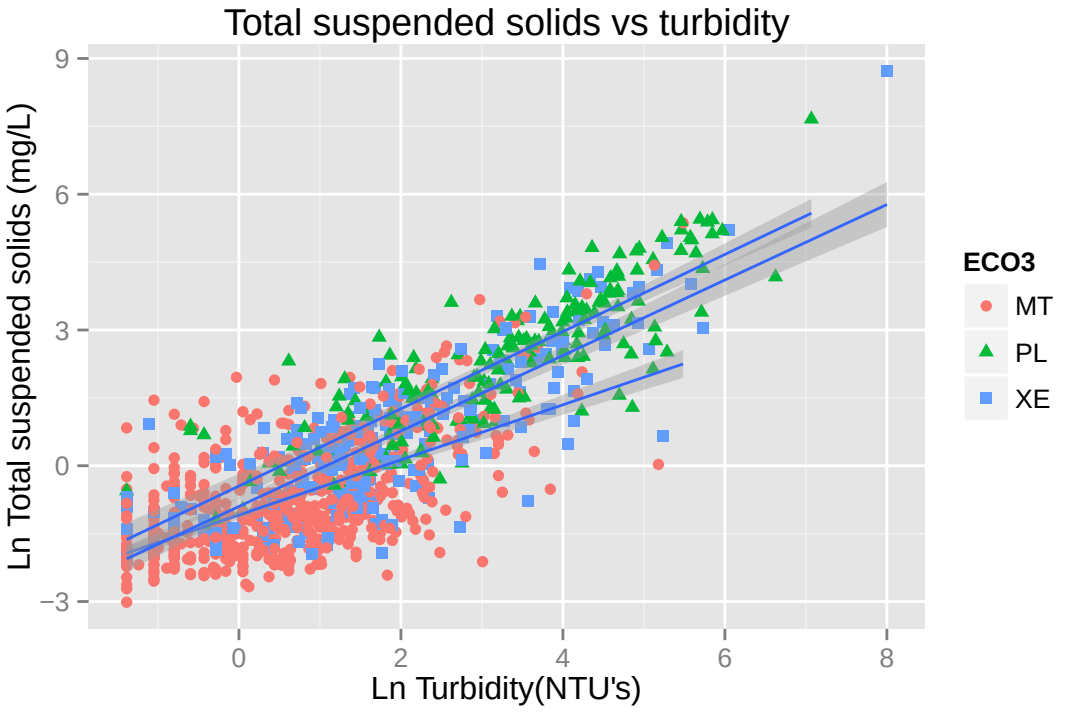

Figure 3: Scatter plots of log transformed total suspended solids vs turbidity for each ecoregion