COMP 333 — Week 3 Basic Plotting

Basic Plotting

Visual descriptions are very important forDescriptive Data Analysis.

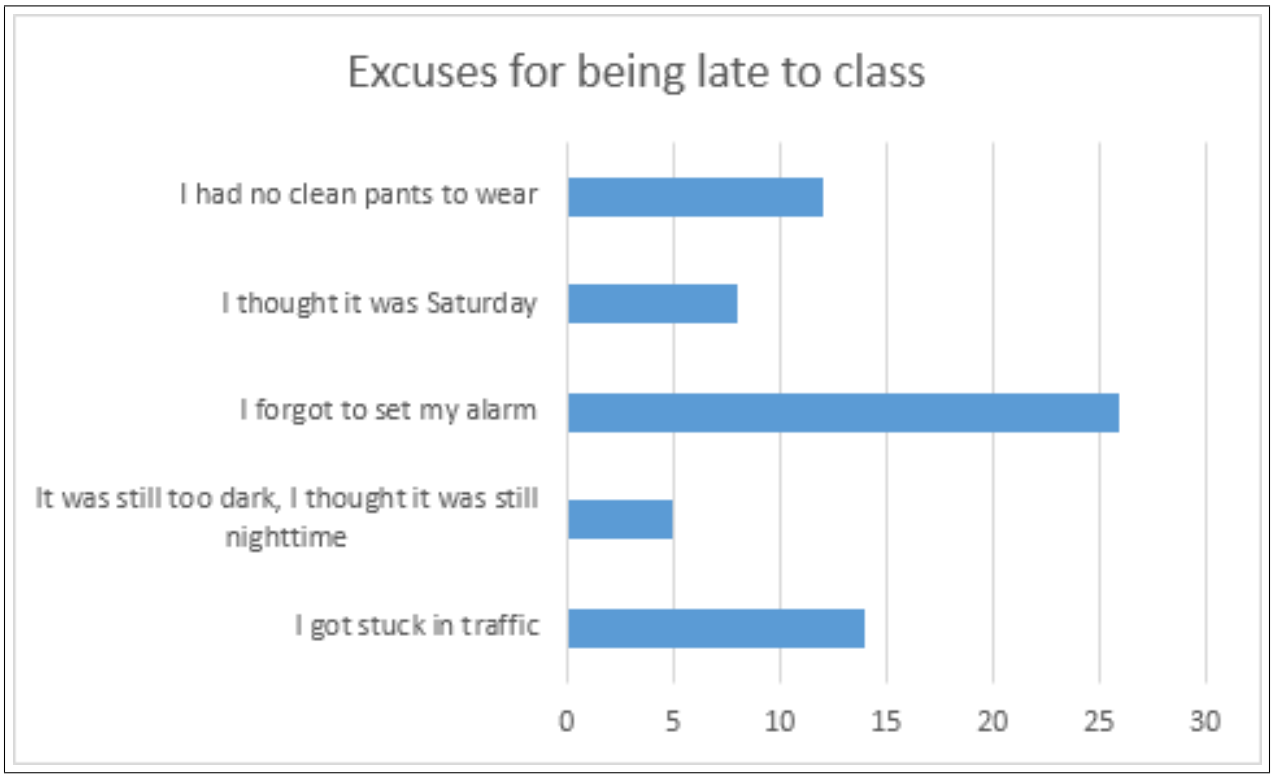
It helps you to understand your data.

Here will will provide an introduction

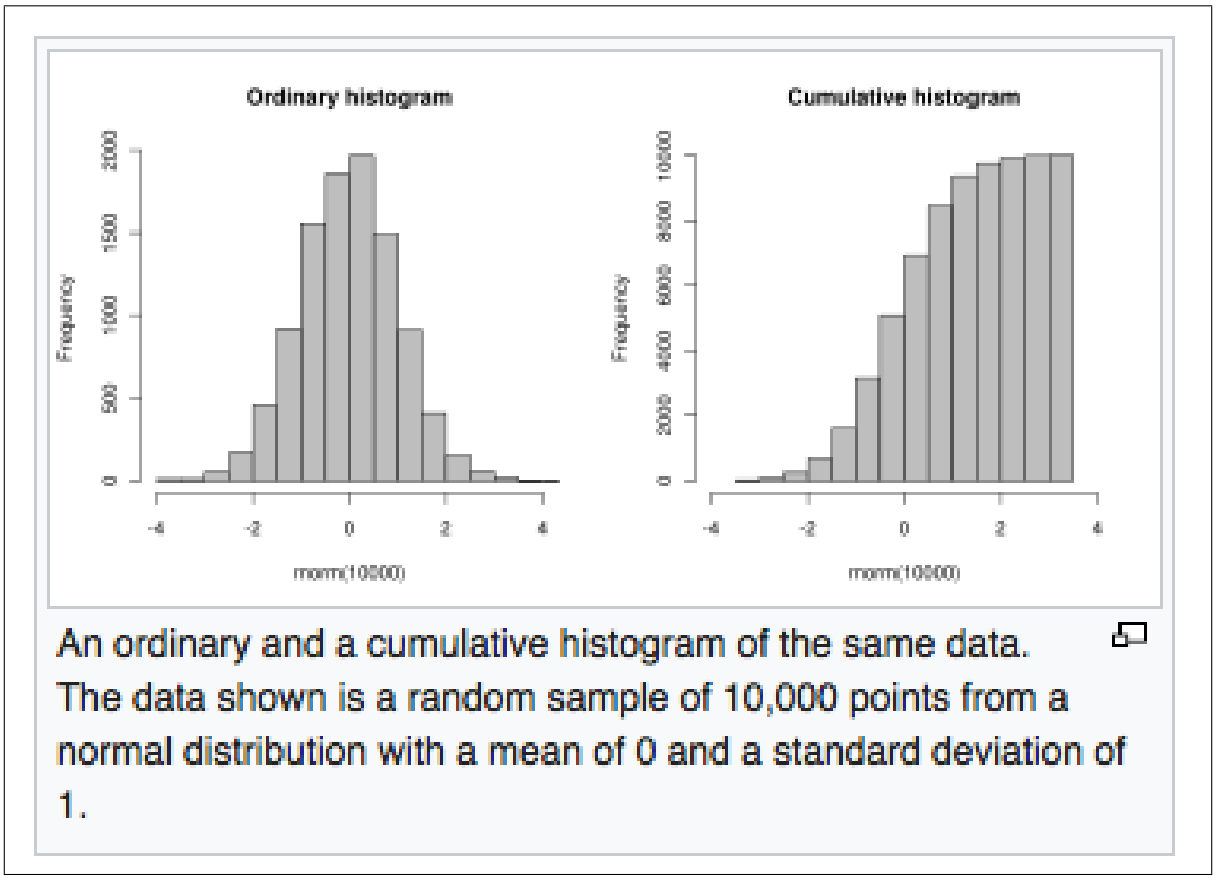
and supplement the material available

- Prof Meyer's video on EDA for the PISA dataset
- The article *An introduction to data visualization in Python*: How to make graphs using `matplotlib`, `pandas` and `seaborn`, by Gilbert Tanner.

Bar Chart

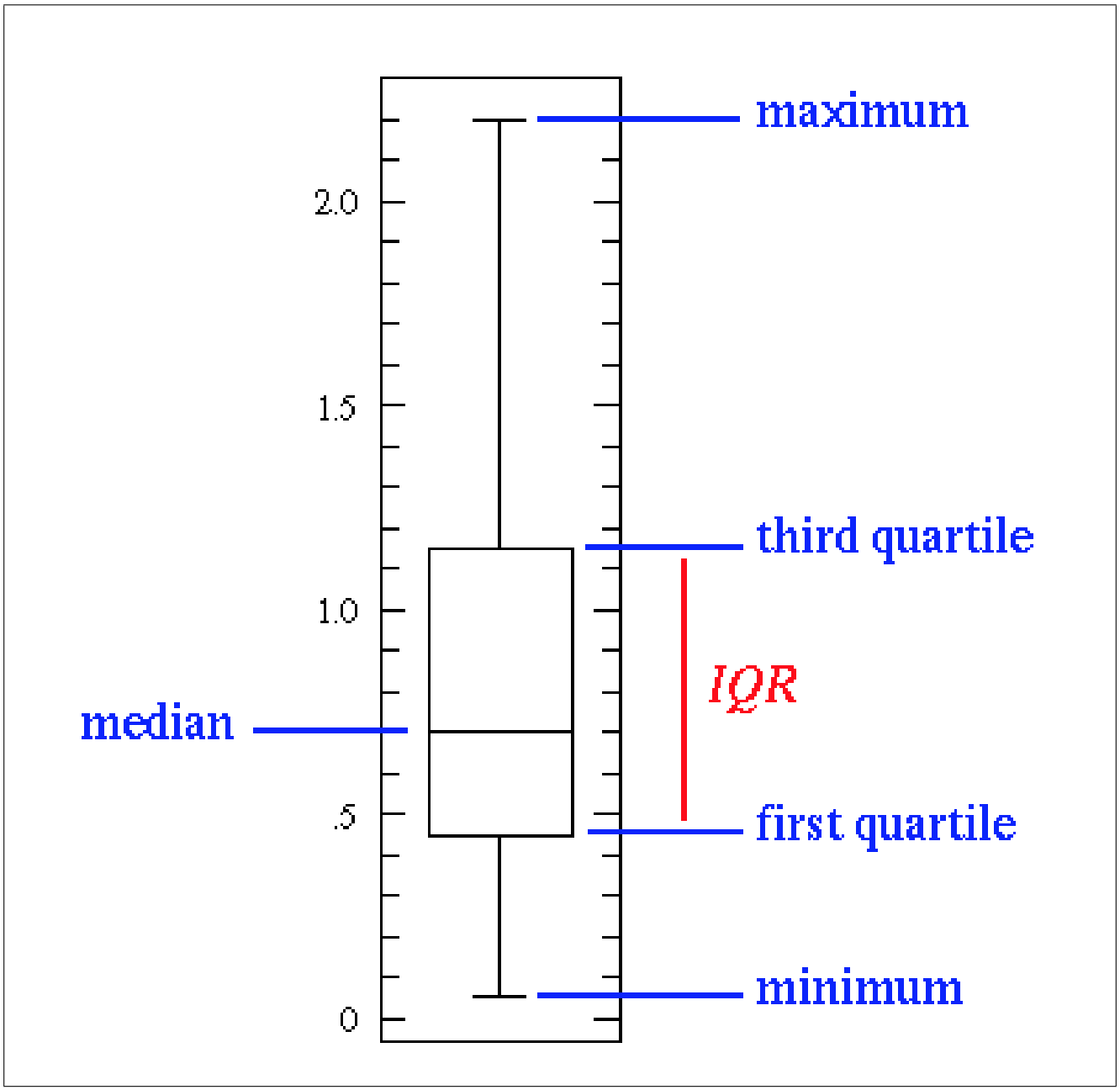

Excuses for being late to class
I had no clean pants to wear
I thought it was Saturday
I forgot to set my alarm
It was still too dark, I thought it was still nighttime
I got stuck in traffic
0
5
10
15
20
25
30

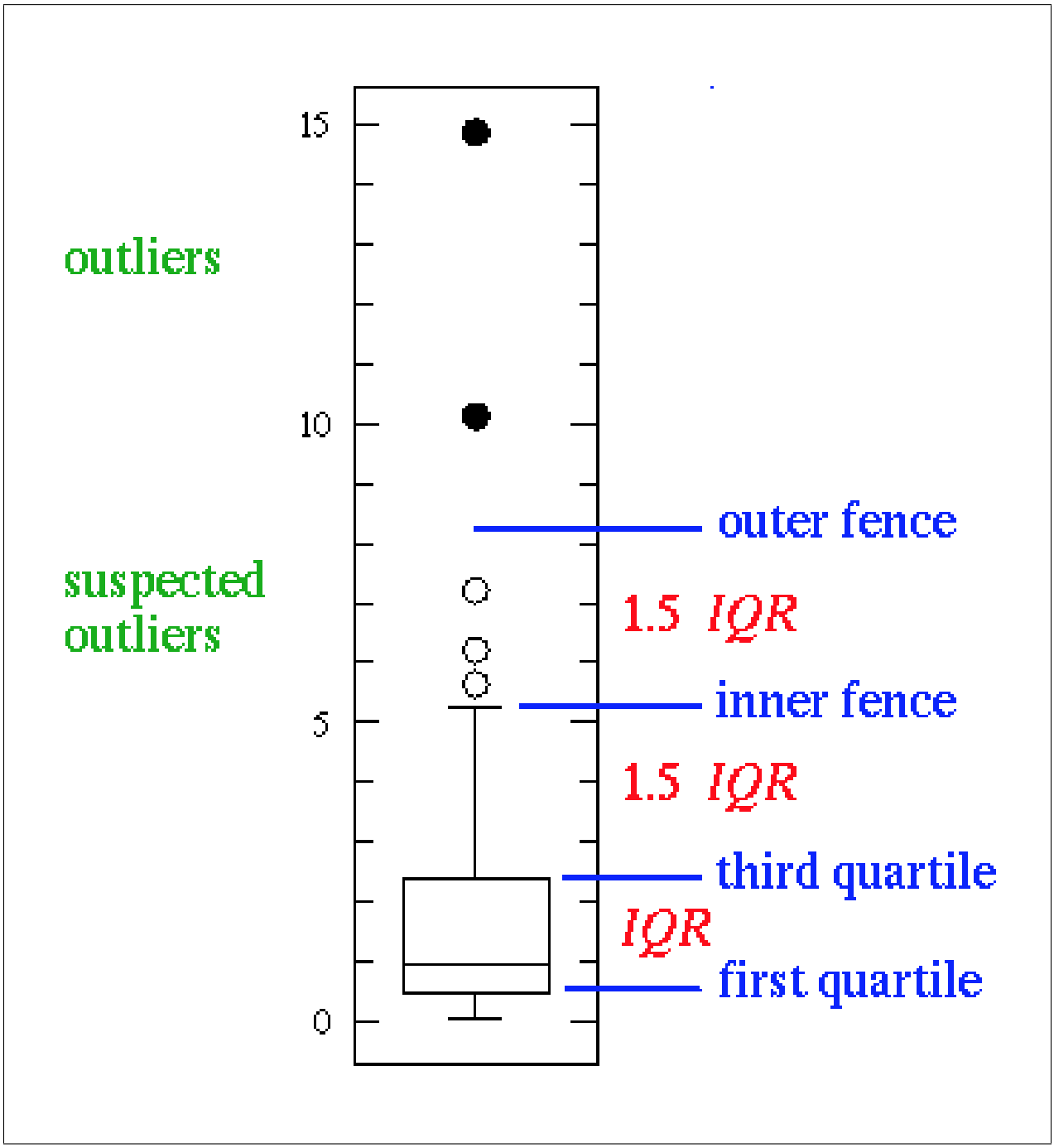
Histogram



An ordinary and a cumulative histogram of the same data. The data shown is a random sample of 10,000 points from a normal distribution with a mean of 0 and a standard deviation of 1.

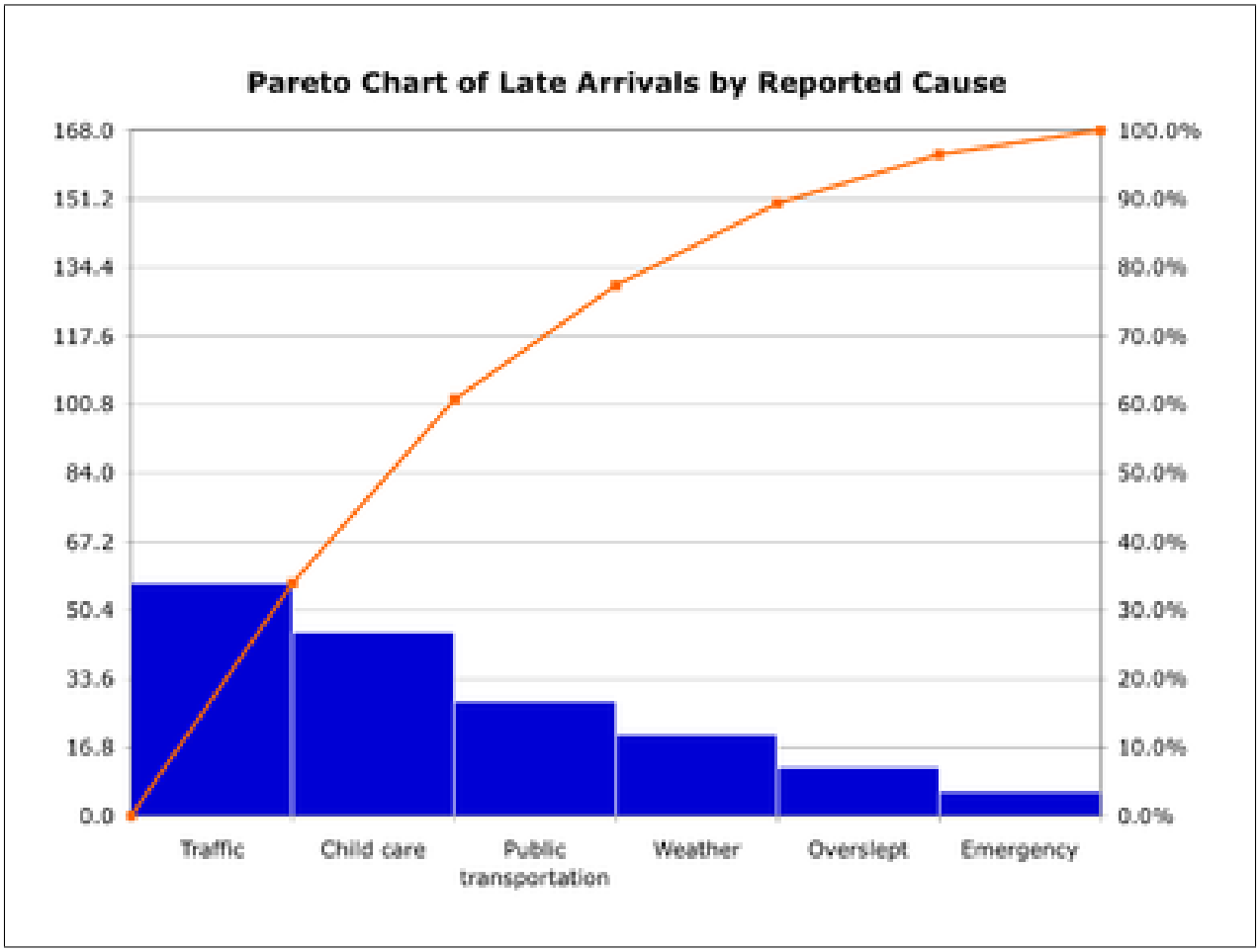
Boxplot


maximum
2.0
1.5
third quartile
1.0
IQR
median
.5
first quartile
minimum
0


15
10
5
0
outliers
suspected outliers
outer fence
1.5 IQR
inner fence
1.5 IQR
third quartile
IQR
first quartile

Pareto Diagram

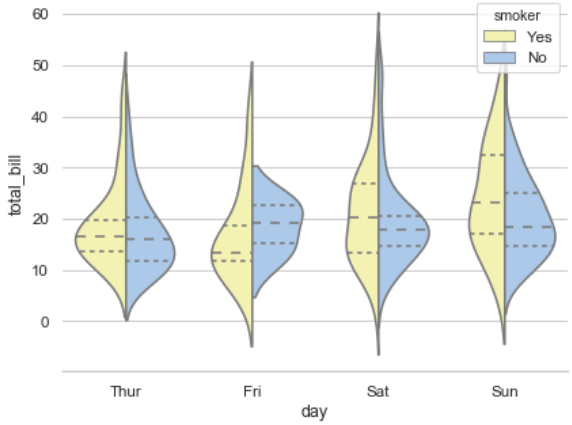
Order by decreasing frequency



Violin Plot

shows frequency too

Grouped violinplots with split violins



Python source code: `[download source: grouped_violinplots.py]`

```
import seaborn as sns
sns.set(style="whitegrid", palette="pastel", color_codes=True)

# Load the example tips dataset
tips = sns.load_dataset("tips")

# Draw a nested violinplot and split the violins for easier comparison
sns.violinplot(x="day", y="total_bill", hue="smoker",
               split=True, inner="quart",
               palette={"Yes": "y", "No": "b"},
               data=tips)
sns.despine(left=True)
```

Comparing Two Attributes

Adapted from Frank E. Harrell Jr. on graphics:

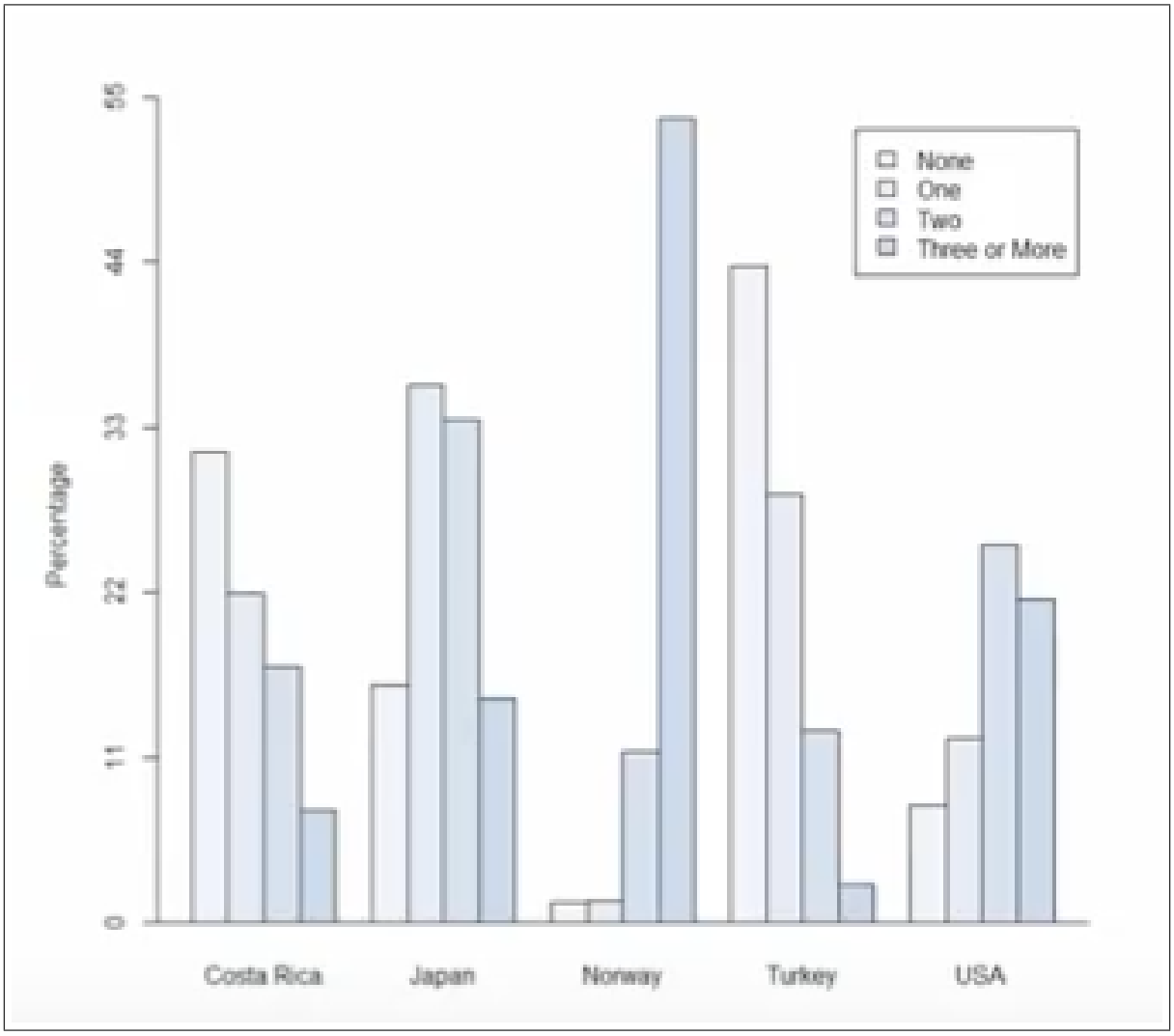
http://biostat.mc.vanderbiltedu/twiki/pub/Main/StatGraphCourse/graphscourse.pdf

Two categorical variables

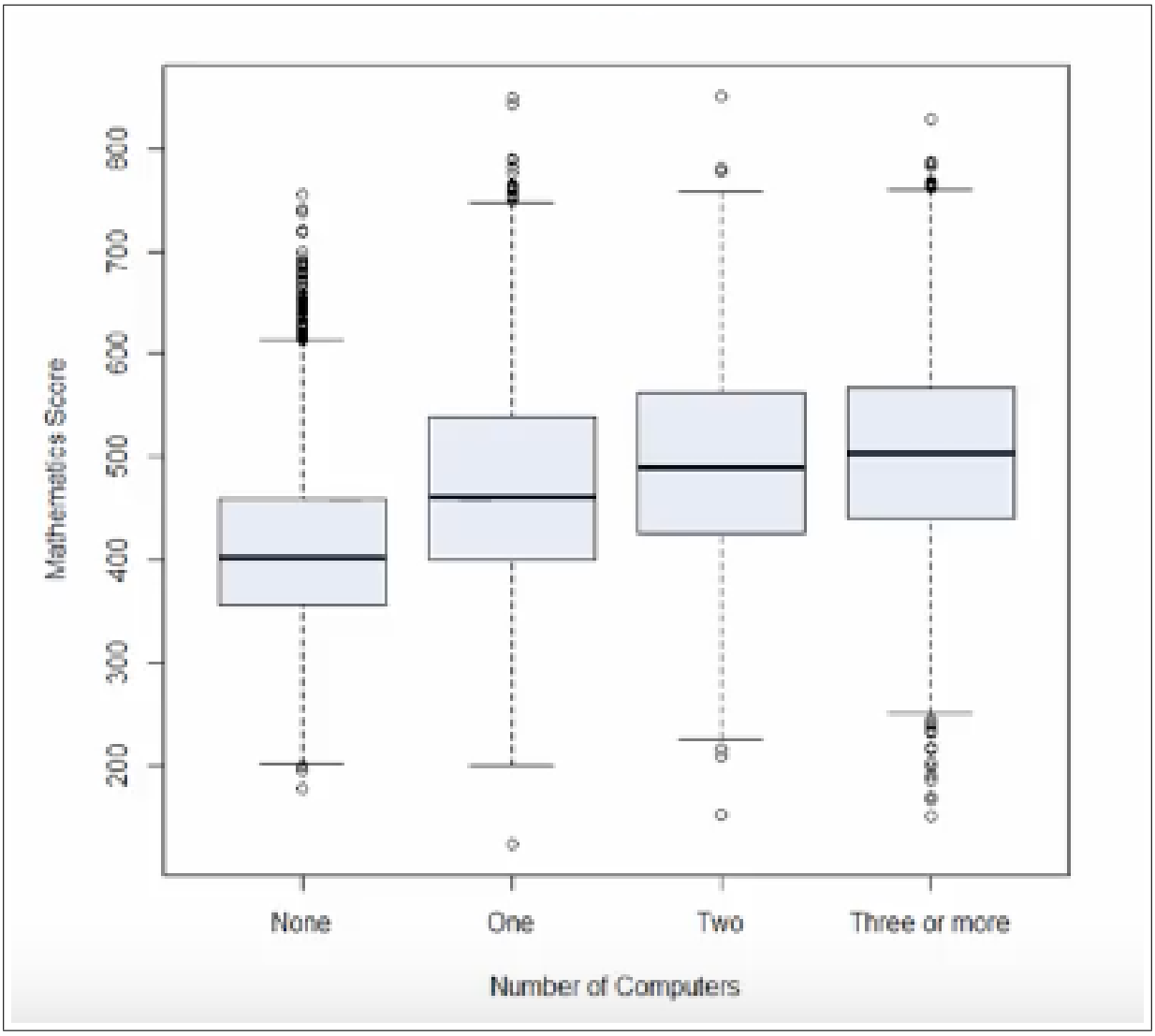
- Use frequency table
  - One categorical variable and other continuous variable
- Box plots of continuous variable values for each category of categorical variable
- Side-by-side dot plots (means + measure of uncertainty, SE or confidence interval)
  - Do not link means across categories!

Two continuous variables

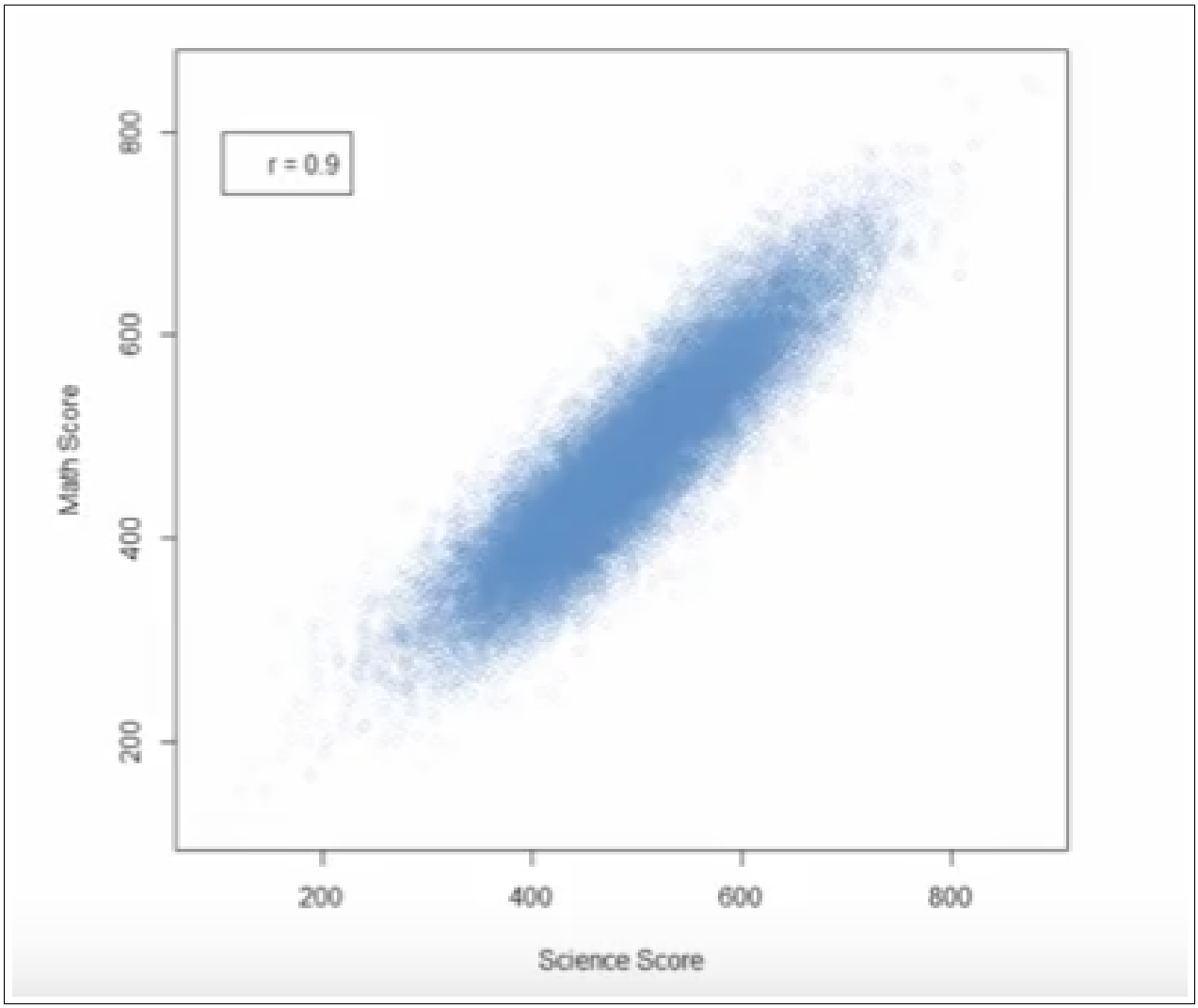
- Scatter plot of raw data if sample size is not too large
- Prediction with confidence bands



Compare categorical and categorical



Compare categorical and continuous



Compare continuous and continuous