Lecture 21

Graphics/Visualizing Data

Manju M. Johny

STAT 330 - Iowa State University

Graphics

Visualizing Data

(mean, median, s, etc)

- Besides reporting numerical summaries to describe data, we can also provide graphical descriptions.
- The most common visualizations for numerical data are:
  1. Histograms
  2. Boxplots
  3. Scatterplots

helps us understand the data quickly

Histograms

Histograms

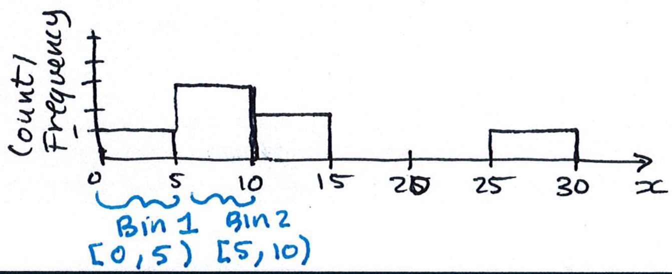
Histograms:

Weight (lbs)
110
145
180
165
⋮

- Most common visualization for one numerical variable
- Can be used to identify potential outliers and anomalies by looking for major "gaps" in histogram

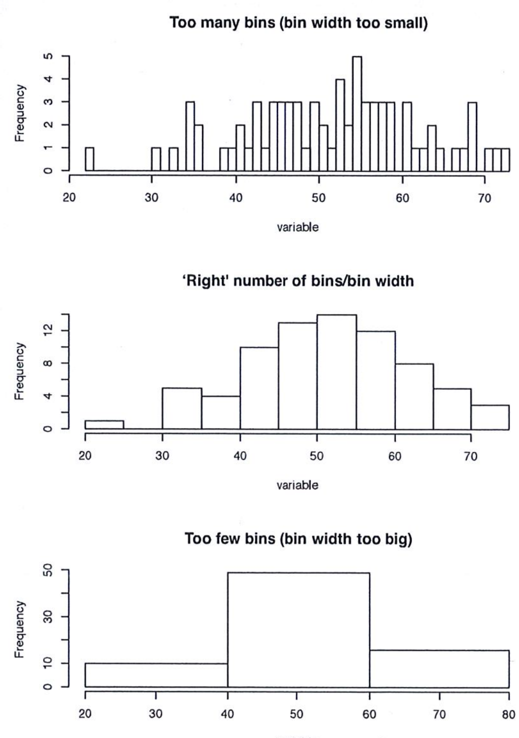
Construction:

1. Start with a data set $x_1, x_2, \ldots, x_n$
2. Divide the data into m intervals (usually of the same width) called "bins": $B_1, B_2, \ldots, B_m$
3. Count how many x's fall into each bin.
4. Draw bars up to the above counts for each bin interval.



Number of Bins

\# of Bins/ Bin width can affect how your histogram looks

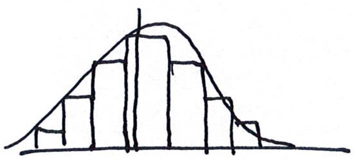


"oversmoothing"

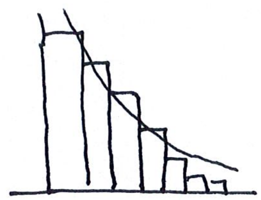
"undersmoothing"

Histograms Cont.

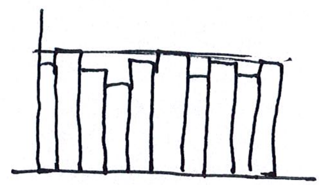
- In the descriptive setting, histograms helps us understand where the data falls
- In the inferential setting, histograms can help us learn about the shape of the probability distribution that generated the data



Normal Distribution?



Exponential Distribution?



Uniform Distribution?

Histogram Cont.

- To understand the shape of the probability distribution, it's useful to use scaled/probability histogram
  - total area under histogram = 1
  - obtained by scaling the height of the histogram
- The Area of the i^{th} Bin (B_i) is ...
  - $\text{Area}_i = \text{height} \cdot \text{width of } B_i$
  - $\text{Area}_i = \frac{\#\text{ of } x\text{'s in } B_i}{n}$

  Then, height of $B_i = \frac{\#\text{ of } x\text{'s in } B_i}{n \cdot \text{width of } B_i}$

  This height gives estimate of probability of your x being in the particular bin.

Boxplots

Boxplots

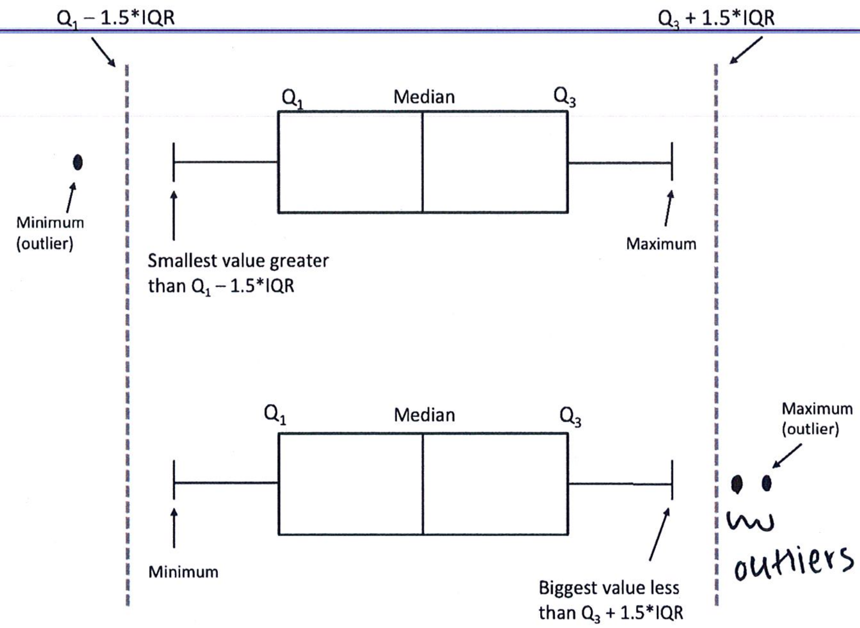
Boxplots:

- Useful for comparing the same numerical variable between multiple groups
- Gives a systematic way to identify outliers

Construction:

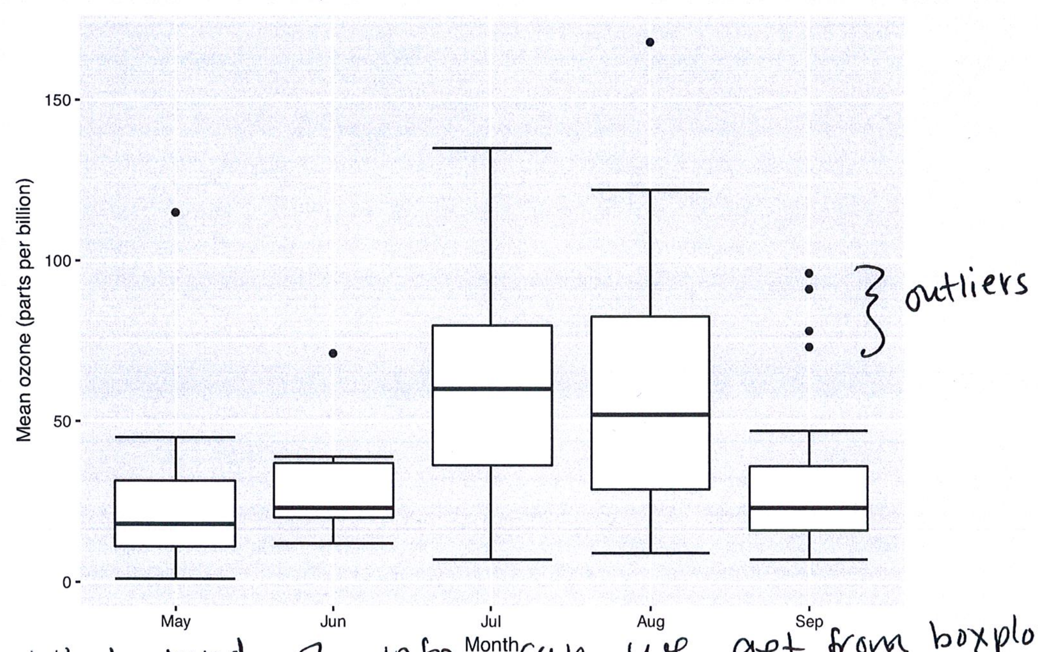
1. 5-point summary: Calculate Min, Q_1, Median, Q_3, Max
2. Box: draw a box between Q_1 and $Q3$, and line at median $IQR = Q_3 - Q_1$
3. Obtain "fences" at $Q_1 - 1.5(IQR)$ and $Q_3 + 1.5(IQR)$.
   $\rightarrow$ box and all non-outlier values are in-between the fences.
4. Whiskers: draw a line from each end of the box out to the closest data value inside the "fence"
5. Outliers: data values outside of the "fences" are represented by dots – these are outliers

Boxplots Cont.



Boxplots Cont.

Can compare variable accross multiple groups easily



What kind of info can we get from boxplots?

- Aug has largest max
- Aug has largest range (max − min)
- July has largest median

Scatterplots

Scatterplots

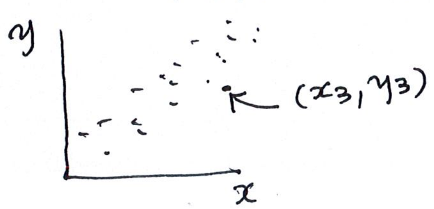
Scatterplots:

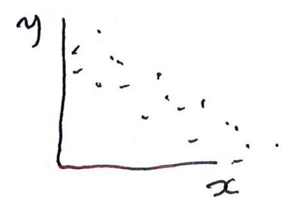
Weight	Height
110	5'1
145	5'5
180	5'8
⋮	⋮

- Used to visualize relationship between 2 numerical variables plotted on (x, y)-plane
  - X = explanatory/predictor variable (x-axis)
  - Y = response/dependent variable (y-axis)
- When the x-axis is time, this is called a time plot (time series)

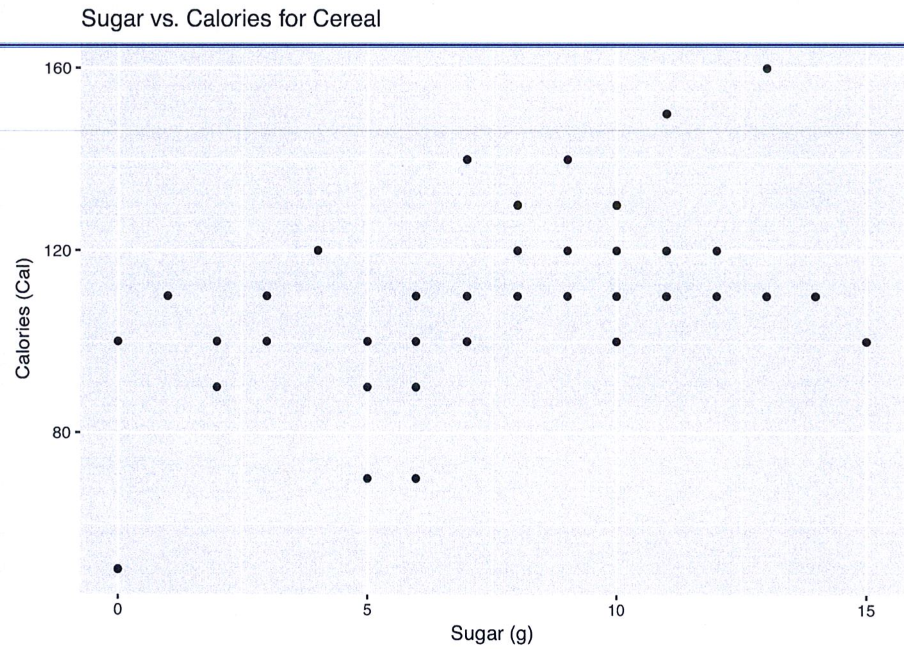
Construction:

1. Obtain x_i and y_i values for each i^{th} subject
2. Arrange into (x, y) pairs: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
3. Plot each (x, y) pair as a point





Scatterplots Cont.

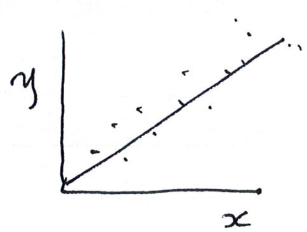


Scatterplots Cont.

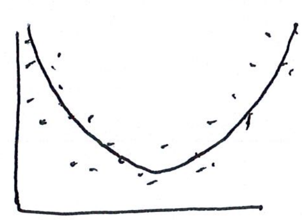
- In the descriptive setting, use scatterplots to understand the general relationship between 2 variables
- In the inferential setting, we develop a model for the relationship between 2 variables of the form:

  $Y = g(X) + \epsilon$

  where $g(\cdot)$ is some function, and ϵ is random error/noise
- Use scatterplots to help learn about the form of $g(\cdot)$



$g(x) = \beta_0 + \beta_1 x$
(linear)



$g(x) = \beta_0 + \beta_1 x + \beta_2 x^2$
(quadratic)