Lecture 19

Descriptive and Graphical Statistics

STAT 330 - Iowa State University

Statistics

Statistics

Definition: Statistics

A *statistic*, $T(X_1, \ldots, X_n)$ is a function of random variables.

- Start with taking a *simple random sample (SRS)* of size n from some population/distribution.
  $X_1, \ldots, X_n \overset{iid}{\sim} f_X(x)$
- We can then obtain *statistics* based on $X_1, \ldots, X_n$ =T(X1,X2,...,Xn)
- Since a statistic is a function $T(\cdot)$ of random variables, the statistic is also a random variable.
- Thus, the statistic will have its own distribution called the *sampling distribution of the statistic* (more on this later!)

Statistics Cont.

Definition: Observed Statistics

The *observed statistics*, $T(x_1, \ldots, x_n)$ is the statistic function with observed values plugged in.

- *Descriptive statistics:* Describing what our sample data looks like (graphically or numerically)
- *Inferential statistics:* Use the statistic to infer/learn about the "true" distribution, $f_X(x)$, that generated the data.

Note:

- Use small letters (x, $\bar{x}$, s^2, etc) to represent observations and observed statistics.
- Use capital letters (X, $\bar{X}$, S^2, etc) to represent random variables.

Mean and Variance

Sample Mean and Variance

Let $X_1, \dots, X_n \overset{iid}{\sim} f_X(x)$ where $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$

- *Sample mean* is defined as $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$
  - $\rightarrow$ estimates the population mean μ.
- *Sample variance* is defined as $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$
  - $\rightarrow$ estimates the population variance σ^2
  - $\rightarrow$ an estimate of the $Var(X) = E[(X - E(X))^2]$ can be found as $\frac{1}{n} \sum_{i=1}^n (X_i - (\bar{X}))^2$
  - $\rightarrow$ typically, n in the above denominator is replaced with $n - 1$ to get S^2 (more on this later)
- *Sample standard deviation* is $S = \sqrt{S^2}$

Note: The quantities above are R.V's since they are functions of R.V's $X_1, \dots, X_n$.

Observed Sample Mean and Variance

- To obtain the *observed sample mean* and *observed sample variance*, plug in observed data values $(x_1, \ldots, x_n)$ into sample mean and variance formulas

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$$

$$s = \sqrt{s^2}$$

Note: The quantities above are not random variables since you have plugged in data values. They are values such as $2.4, 100,$ *etc*.

Quantiles

Quantiles

Definition: Quantiles

The q^{th} *quantile* of a distribution, $f_X(x)$, is a value x such that $P(X < x) \leq q$ and $P(X > x) \leq 1 - q$.

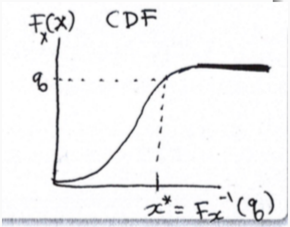
This is also called the $100 \cdot q^{th}$ *percentile*.

$Q_1 = 0.25^{th}$ quantile, $Q_2 = 0.5^{th}$ quantile (median), and $Q_3 = 0.75^{th}$ quantile

Definition: Quantile Function

The *quantile function* is defined as:

$$F_X^{-1}(q) = min\{x : F_X(x) \geq q\}$$



Median

(50% of my observations are less than the median)

The *median* is the 0.5^{th} quantile (or 50^{th} percentile)

$\rightarrow$ can be written as $F_X^{-1}(0.5)$

The *sample median* is calculated by:

$X_{(k)}$ is an "order statistic"
$X_{(1)}$ = min
$X_{(n)}$ = max
$X_{(5)}$ = 5th ordered value

1. Order sampled values in increasing order: : $X_{(1)}, \ldots, X_{(n)}$
   - If n is odd, take the middle value
     $\rightarrow$ median $= X_{(\lceil \frac{n}{2} \rceil)}$

     Ex: 2.2 4.1 7.3
     median = 4.1
   - If n is even, average the two middle values
     $\rightarrow$ median $= \frac{X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}}{2}$

     Ex: 2.2 3 4 7.3
     median = 3.5

Note: Since the above values are functions of R.V's, they are R.Vs.

Obtain the *observed sample median* by plugging in the observed values $(x_1, \ldots, x_n)$ from data.

Q_1 and Q_3

Other sample quantiles we are typically interested in are

- $Q_1 = 0.25^{th}$ quantile
- $Q_3 = 0.75^{th}$ quantile

Many ways to calculate quantiles. Our method for a general q^{th} sample quantile is ...

1. Compute $(n+1) \cdot q$
   - If this value is an integer, use $(n+1)q^{th}$ ordered value
   - Else, use the average of the 2 surrounding values

Example

Example 1: A sample $X_1, \dots, X_n \overset{iid}{\sim} f_X(x)$ was taken where $X_i =$ CPU time for a randomly chosen task. The ordered observed values are $15, 34, 35, 36, 43, 48, 49, 62, 70, 82$ (secs)
The observed $\cdots$

- sample mean: $\bar{x} = \frac{15+34+\cdots+82}{10} = 47.4$
- sample variance:
  $s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} = \frac{(15-47.4)^2+\cdots+(82-47.4)^2}{10-1} = 384.04$
- sample standard dev: $s = \sqrt{s^2} = \sqrt{384.04}$
- sample median: $n = 10 \to$ even (average the two middle vals)
  median $= \frac{x_{(5)}+x_{(6)}}{2} = \frac{43+48}{2} = 45.5$

- sample Q_1: $(n+1)q = (10+1)(0.25) = 2.75$ (take average of 2^{nd} and 3^{rd} ordered values)
  $Q_1 = \frac{x_{(2)}+x_{(3)}}{2} = \frac{34+35}{2} = 34.5$
- sample Q_3 : $(n+1)q = (10+1)(0.75) = 8.25$ (take average of 8^{th} and 9^{th} ordered values)
  $Q_3 = \frac{x_{(8)}+x_{(9)}}{2} = \frac{62+70}{2} = 66$

Right now, we're only using these statistics to describe the sample of CPU speeds.

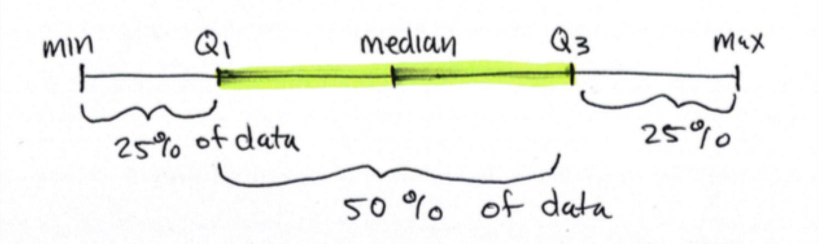
- sample mean and median (Q_2) tell us "typical" values
- sample variance tells us how "spread out" / how variable the data are
- Q_1 and Q_3 "rank" where values fall in our sample

Mode, Range, IQR

Mode, Range, and IQR

Other common descriptive statistics to describe the data:

- *Mode:* The most frequent value in our sample. Can have multiple modes in data set
- *Range:* Max - Min $= X_{(n)} - X_{(1)}$
  $\rightarrow$ describes the "total" variability of the data
- *Interquartile Range (IQR):* $Q_3 - Q_1$
  $\rightarrow$ describes the variability of the middle 50% of data



Robust Statistics

- With all the different options for statistics, how do we choose which ones to use?

  $\rightarrow$ It depends on your data set
- Statistics that are not affected by extreme values are called *robust statistics*

  Robust – median, IQR

  Not Robust – mean, variance, s, range

Example 2:

Imagine your favorite celebrity (who's extremely rich!) moves in next door. What happens to the statistics after they move in?

statistic	before celeb	after celeb	robust?
mean	40K	way bigger	no
median	40K	$\approx$ same	yes
std dev	10K	way bigger	no

Graphical Statistics

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:

helps us understand the data quickly

  1. Histograms
  2. Boxplots
  3. Scatterplots

Histograms

Histograms

Weight (lbs)

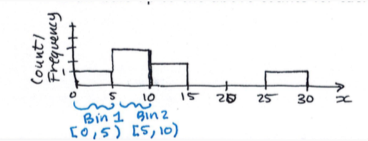
145
110
180
etc

Histograms:

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

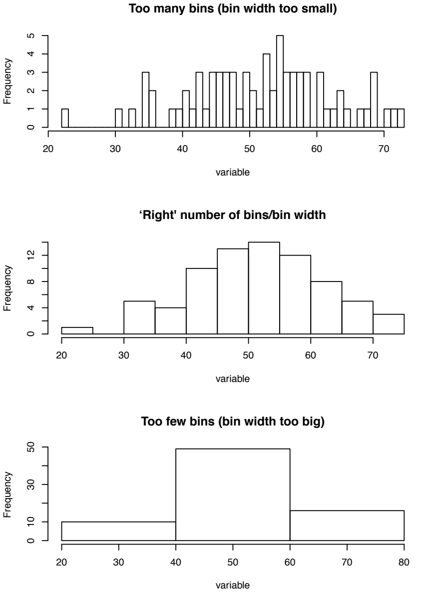
Construction:

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

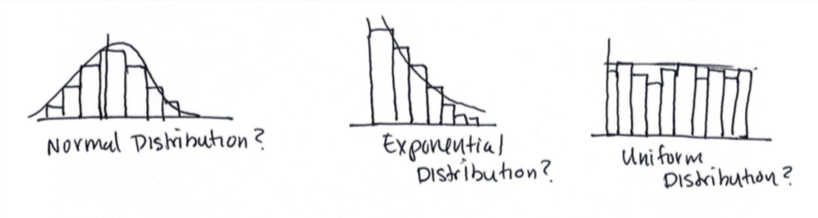


"undersmoothing"

"oversmoothing"

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



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:

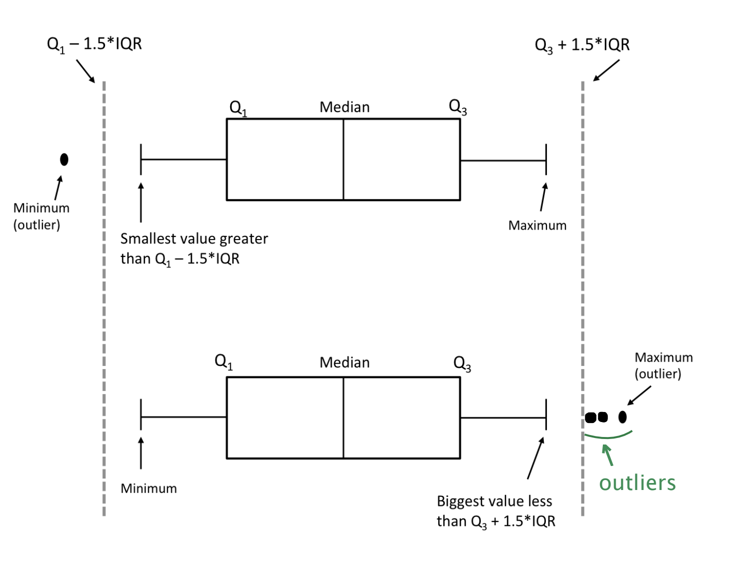
- Useful for comparing the same numerical variable between multiple groups
- Gives a systematic way to identify outliers

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
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

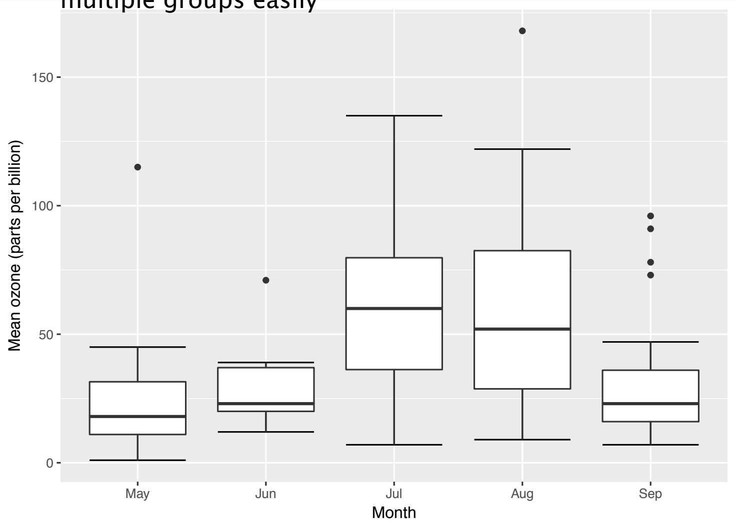
IQR = Q3 – Q1

Boxplots Cont.



Boxplots Cont.

can compare variable across multiple groups easily



Scatterplots

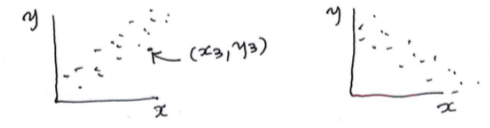
Scatterplots

Scatterplots:

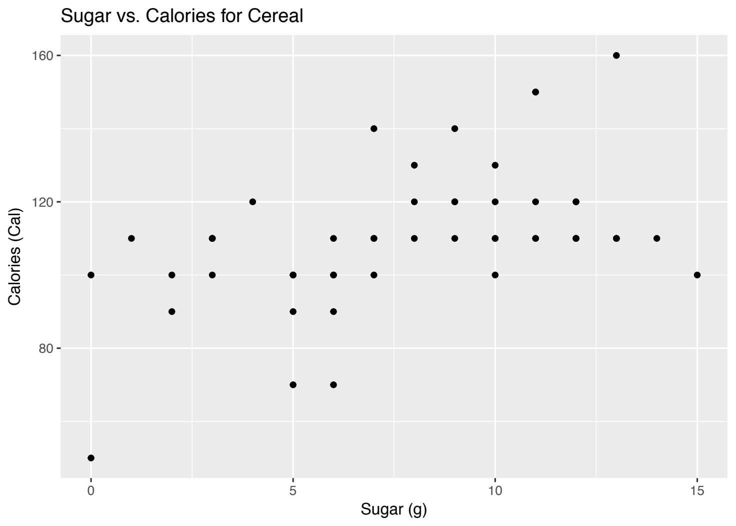
- Used to visualize relationship between 2 numerical variables plotted on (x, y)-plane
  - X = explanatory/predictor variable (x-axis)
  - Y = response/dependent variable (y-axis)
- When the x-axis is time, this is called a time plot (time series)

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 $g(\cdot)$