Week 2!!!

Learning Unit 2 – Graphical Summaries

Your Plotting Professor Colton

LU 2 Outline

Introduction

- Why Summaries and Plots?

Frequency Tables

- Frequency Tables

Frequency Table Graphs

- Histograms
- Relative Frequency Histograms

Other Visual Tools

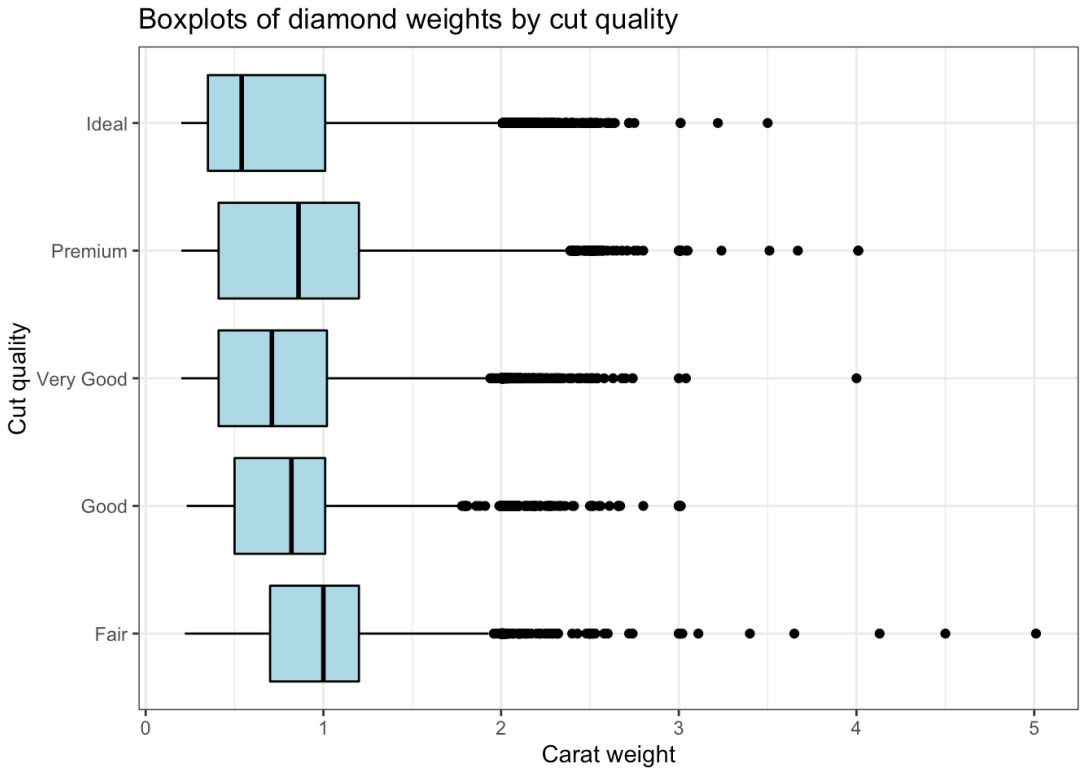
- Stem and Leaf Plots
- Dot Plots
- Time Series Plots
- Measures of Shape

Introduction – Why Summaries and Plots?

- Terrible idea to display raw information! Way too much!
- Two parts of any data analysis: Understanding and Relating!!

```
  [1] Very Good Ideal     Very Good Very Good Fair      Premium   Ideal
  [8] Very Good Very Good Ideal     Ideal     Fair      Very Good Very Good
 [15] Ideal     Premium   Ideal     Good      Ideal     Ideal     Very Good
 [22] Premium   Ideal     Ideal     Ideal     Ideal     Ideal     Ideal
 [29] Ideal     Fair      Premium   Premium   Premium   Good      Premium
 [36] Premium   Very Good Ideal     Very Good Premium   Ideal     Ideal
 [43] Ideal     Very Good Very Good Premium   Very Good Ideal     Premium
 [50] Very Good Very Good Very Good Ideal     Very Good Very Good Premium
 [57] Ideal     Fair      Good      Premium   Very Good Very Good Ideal
 [64] Ideal     Ideal     Ideal     Ideal     Premium   Ideal     Ideal
 [71] Fair      Very Good Premium   Ideal     Ideal     Ideal     Very Good
 [78] Premium   Premium   Ideal     Very Good Premium   Very Good Very Good
 [85] Premium   Ideal     Ideal     Ideal     Good      Very Good Ideal
 [92] Ideal     Ideal     Ideal     Ideal     Very Good Ideal     Premium
 [99] Ideal     Ideal
Levels: Fair < Good < Very Good < Premium < Ideal
```

```
  [1] 0.35 0.30 1.25 0.30 0.30 1.51 0.57 0.73 1.01 0.54 0.42 2.01 1.71 2.01 0.51
 [16] 0.31 0.35 0.70 0.90 0.90 2.48 0.74 1.12 0.51 0.86 2.01 0.50 0.41 0.38 0.90
 [31] 1.02 1.13 2.01 1.00 1.50 0.73 1.34 1.01 1.20 1.36 1.04 1.27 0.30 0.23 1.01
 [46] 0.40 1.54 0.40 1.52 0.93 1.02 1.03 0.30 0.41 0.31 0.76 0.55 0.58 0.61 0.42
 [61] 0.91 1.00 0.70 0.71 0.31 0.70 0.57 0.42 0.56 0.32 0.70 0.60 0.40 0.32 0.30
 [76] 0.50 0.43 0.35 1.74 1.02 1.24 0.31 1.07 1.50 0.33 0.42 0.31 2.50 1.23 1.25
 [91] 0.31 0.34 1.14 1.52 0.60 1.06 0.90 1.14 0.41 1.09
```



Review + New

- What type of visual we can make depends on what type of data we have.

Qualitative (Categorical) Data

- Non-Numerical data with different categories.
- Ex) States, letter grades, class standing, etc.

Quantitative Data

- Numerical data, counts or measurements
- Arithmetic operations such as adding and averaging make sense
- Ex) Income, GPA, Height, Weight, etc.

- In creating graphs, it is important to first consider the **distribution** of a variable.
    - The **distribution** of a variable is a list of what values the variable can take on as well as how often it takes on these values.
- Often, the distribution of quantitative variables is summarized in a table like the following example.

Frequency Tables for Numeric Data

Frequency Tables

- Used to organize both **quantitative** and **qualitative** data sets.
- They also are helpful in some generating graphs we will discuss later!

How to Construct for Numeric Data

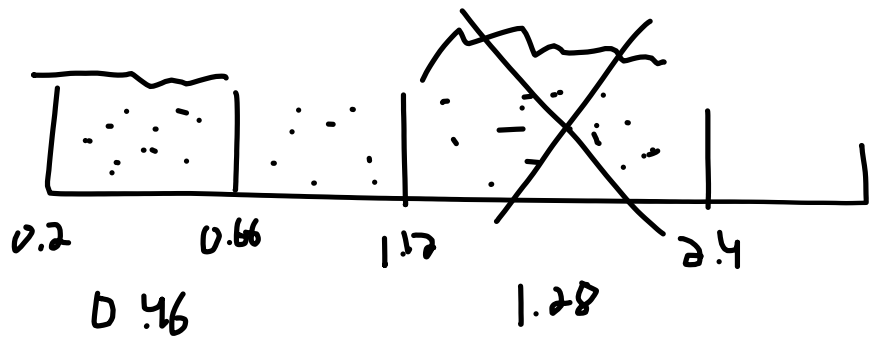
1. Data is **binned,** i.e. grouped together. All bins (aka **classes**) will have equal length.
   - This length is referred to as the **bin / class width = upper limit – lower limit**
2. Then we *count* the number of observations in each bin.

- This count is referred to as the **Frequency**.
- Can also find the **Relative Frequency**, which is just the *proportion (or percent)* of data in each bin.
  - This is just the **Frequency** divided by the **Total**.

: count / sample size

Carat	Frequency	Relative Frequency
[0.2,0.66]	95 / 200	0.475
(0.66,1.12]	56 / 200	0.280
(1.12,1.58]	36	0.180
(1.58,2.04]	11	0.055
(2.04,2.5]	2	0.010
	Total: 200	1.00

* This can be referred to as **grouped data** now



LCQ: Frequency Tables

- Data (20 observations):
  - 38, 33, 5, 5, 47, 29, 24, 42, 3, 18,
    30, 46, 25, 44, 40, 42, 39, 44, 29, 13

Construct a Frequency Table

Bin	**Frequency**	**Relative Frequency**
0-10		
10-20		
20-30		
30-40		
40-50		
Total:		

LCQ: Frequency Tables

- Data (20 observations):
  - 38, 33, 5, 5, 47, 29, 24, 42, 3, 18,
    30, 46, 25, 44, 40, 42, 39, 44, 29, 13

Construct a Frequency Table

Bin	Frequency	Relative Frequency
0-10	*3*	*3/20 = 0.15*
10-20	*2*	*2/20 = 0.1*
20-30	*4*	*0.20*
30-40	*4*	*0.20*
40-50	*7*	*0.35*
Total:	*20*	*1.00*

$20 \le x < 30$

$30 \le x < 40$

30

Check work:

3 + 2 + 4 + 4 + 7 ==== Total 20

15% + 10% + 20% + 20% + 35% ==== 100%

Frequency Tables

Choosing the Bins

- Subjective choice
  - Depends on how granular (**closely**, exactly) you want to show the data.

- With more bins, each bin gets **smaller** (smaller class width).
  - So the data gets more spread out across the bins (smaller frequencies and relative frequencies).

Carat	Frequency	Relative Frequency
[0.2,0.66]	95	0.475
(0.66,1.12]	56	0.280
(1.12,1.58]	36	0.180
(1.58,2.04]	11	0.055
(2.04,2.5]	2	0.010

Carat	Frequency	Relative Frequency
[0.2,0.487]	70	0.350
(0.487,0.775]	43	0.215
(0.775,1.06]	31	0.155
(1.06,1.35]	28	0.140
(1.35,1.64]	15	0.075
(1.64,1.92]	4	0.020
(1.92,2.21]	7	0.035
(2.21,2.5]	2	0.010

Frequency Tables

```
 [1] 0.35 0.30 1.25 0.30 0.30 1.51 0.57 0.73 1.01 0.54 0.42 2.01 1.71 2.01 0.51
[16] 0.31 0.35 0.70 0.90 0.90 2.48 0.74 1.12 0.51 0.86 2.01 0.50 0.41 0.38 0.90
[31] 1.02 1.13 2.01 1.00 1.50 0.73 1.34 1.01 1.20 1.36 1.04 1.27 0.30 0.23 1.01
[46] 0.40 1.54 0.40 1.52 0.93 1.02 1.03 0.30 0.41 0.31 0.76 0.55 0.58 0.61 0.42
[61] 0.91 1.00 0.70 0.71 0.31 0.70 0.57 0.42 0.56 0.32 0.70 0.60 0.40 0.32 0.30
[76] 0.50 0.43 0.35 1.74 1.02 1.24 0.31 1.07 1.50 0.33 0.42 0.31 2.50 1.23 1.25
[91] 0.31 0.34 1.14 1.52 0.60 1.06 0.90 1.14 0.41 1.09
```

Advantage of Using Frequency Tables

- Great for **condensing** and **summarizing** the raw data.

Disadvantage of Using Frequency Tables

- We **lose** information.
- We **no longer** know what the specific values were, only what bin (range of values) they are in.

- So why not just use a lot of bins????
  - Then we would be keeping more information right??

Carat	Frequency	Relative Frequency
[0.2,0.66]	95	0.475
(0.66,1.12]	56	0.280
(1.12,1.58]	36	0.180
(1.58,2.04]	11	0.055
(2.04,2.5]	2	0.010

Frequency Tables

Balancing Act

- Yes, using more bins keeps more info…
- But then it loses the conciseness that made it a good representation in the first place.

- Hard to really get anything from this frequency table…

- So you have to be smart about choosing the number / width of the bins (classes)!
  - We will learn how to set the ranges for each bin next class!

Carat	Frequency	Relative Frequency
[0.2,0.315]	30	0.150
(0.315,0.43]	40	0.200
(0.43,0.545]	10	0.050
(0.545,0.66]	15	0.075
(0.66,0.775]	18	0.090
(0.775,0.89]	2	0.010
(0.89,1]	13	0.065
(1,1.12]	23	0.115
(1.12,1.23]	10	0.050
(1.23,1.35]	11	0.055
(1.35,1.46]	1	0.005
(1.46,1.58]	14	0.070
(1.58,1.69]	1	0.005
(1.69,1.81]	3	0.015
(1.92,2.04]	7	0.035
(2.38,2.5]	2	0.010

Cumulative Frequency

Cumulative Frequency

- This is a running total of the frequencies
  - So for a class, it is the number of values less than or equal to that class.
- Can also do this as **cumulative relative frequency**.
  - So just continually add the relative frequencies as you move down the classes

Carat	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Relative Frequency
[0.2,0.66]	95	0.475	95	0.475
(0.66,1.12]	56	0.280	151	0.755
(1.12,1.58]	36	0.180	187	0.935
(1.58,2.04]	11	0.055	198	0.990
(2.04,2.5]	2	0.010	200	1.000

LCQ: Cumulative Frequency

- Data (20 observations):
  - 38, 33, 5, 5, 47, 29, 24, 42, 3, 18, 30, 46, 25, 44, 40, 42, 39, 44, 29, 13

Construct the rest of the earlier Frequency Table

Bin	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Relative Frequency
0-10	*3*	*0.15*		
10-20	*2*	*0.1*		
20-30	*4*	*0.20*		
30-40	*4*	*0.20*		
40-50	*7*	*0.35*		
Total:	*20*	*1*		

LCQ: Cumulative Frequency

- Data (20 observations):
  - 38, 33, 5, 5, 47, 29, 24, 42, 3, 18, 30, 46, 25, 44, 40, 42, 39, 44, 29, 13

Construct the rest of the earlier Frequency Table

Bin	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Relative Frequency
0-10	*3*	*0.15*	*3*	*3 / 20 = 0.15*
10-20	*2*	*0.1*	*3 + 2 = 5*	*5 / 20 = 0.25*
20-30	*4*	*0.20*	*5 + 4 = 9*	*0.45*
30-40	*4*	*0.20*	*13*	*0.65*
40-50	*7*	*0.35*	*20*	*1*
Total:	*20*	*1*		

Frequency Table Graphs

Graphs (in general)

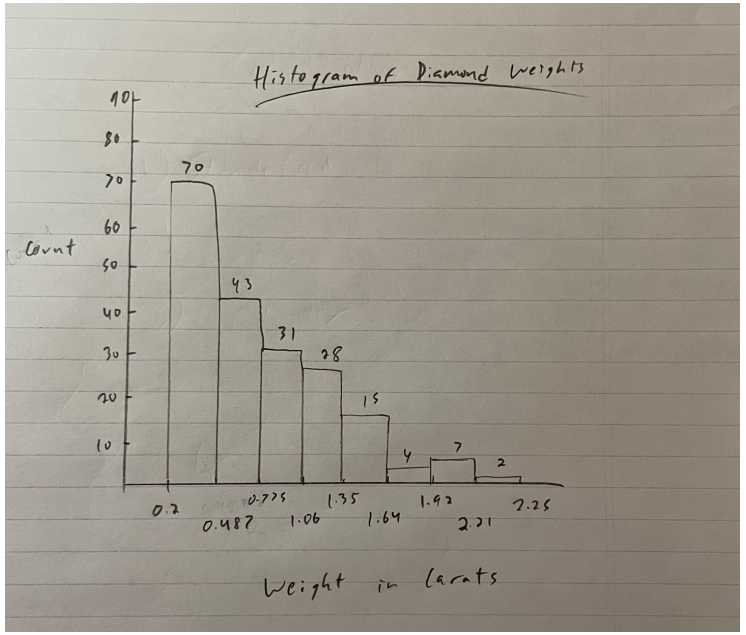
- Use graphs to display data visually.
- Important to know which type of data (qualitative or **quantitative**) each different type graph we discuss shows.

Frequency Table Graphs

- Frequency tables help construct Histograms and Relative Frequency Histograms.
- To create these in graphs by hand or in Excel, you first have to make a frequency table.

Histograms

Carat	Frequency	Relative Frequency
[0.2,0.487]	70	0.350
(0.487,0.775]	43	0.215
(0.775,1.06]	31	0.155
(1.06,1.35]	28	0.140
(1.35,1.64]	15	0.075
(1.64,1.92]	4	0.020
(1.92,2.21]	7	0.035
(2.21,2.5]	2	0.010



Histograms

- For **Quantitative** data.
- Each bar represents a bin from the Freq table.

Features

- **Bars touch**! This is because we have to account for every possible value along the x-axis.
- Each bar is the **same width**.
  - This makes comparisons across bars valid!
- Heights of each bar represents the Frequency for that bin.
- Sum of all the heights equals the total sample size.
- **CAN'T tell the specific value of observations, only the interval its in.**
  - This is why we lose information when grouping the data for Freq Tables / Histograms.

Advantages

- Pretty simple.
- Shows **shape** and **mode** (the most common value) well!

LCQ: Histograms

- Data (20 observations):
  - 38, 33, 5, 5, 47, 29, 24, 42, 3, 18, 30, 46, 25, 44, 40, 42, 39, 44, 29, 13

Sketch a Histogram

Bin	Frequency	Relative Frequency
0-10	*3*	*3/20 = 0.15*
10-20	*2*	*2/20 = 0.1*
20-30	*4*	*0.2*
30-40	*4*	*0.2*
40-50	*7*	*0.35*
Total:	*20*	*1*

Questions we can ask:

Context: Lets say this data represents the scores for a football team.

1) What is the range of scores?

2) How many games had 30 or more points (score ≥ 30)?

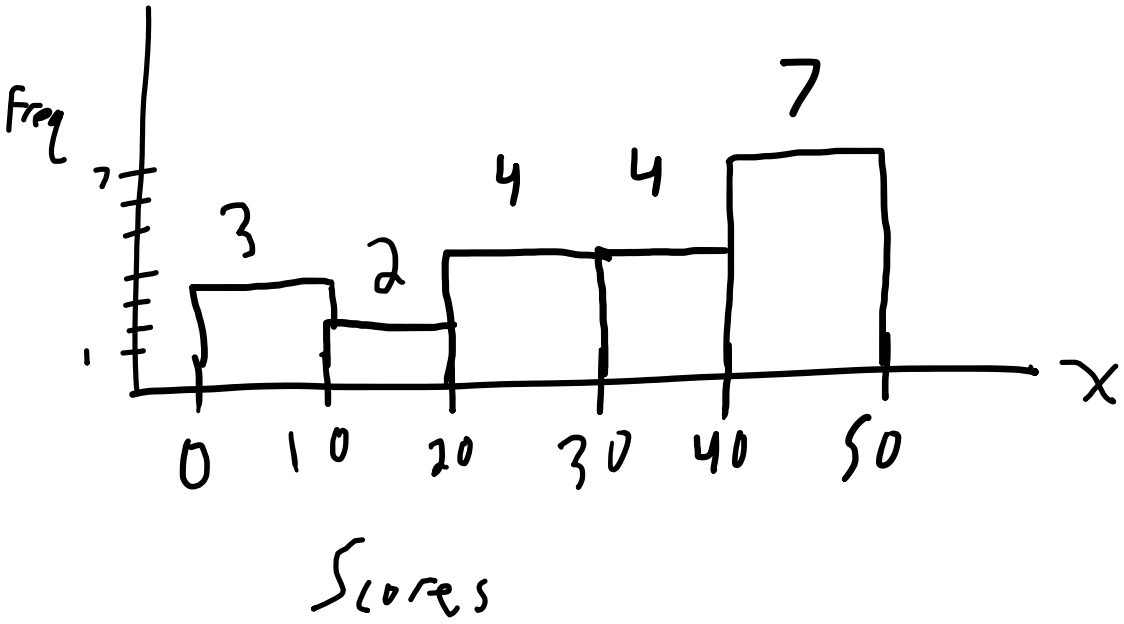
3) What percentage of games had less than 20 points (score < 20)?

LCQ: Histograms

- Data (20 observations):
  - 38, 33, 5, 5, 47, 29, 24, 42, 3, 18, 30, 46, 25, 44, 40, 42, 39, 44, 29, 13

Bin	Frequency	Relative Frequency
0-10	*3*	*3/20 = 0.15*
10-20	*2*	*2/20 = 0.1*
20-30	*4*	*0.2*
30-40	*4*	*0.2*
40-50	*7*	*0.35*
Total:	*20*	*1*

Sketch a Histogram



Questions we can ask:

Context: Lets say this data represents the scores for a football team.

1) What is the range of scores? *Scores range from 0 to 50*

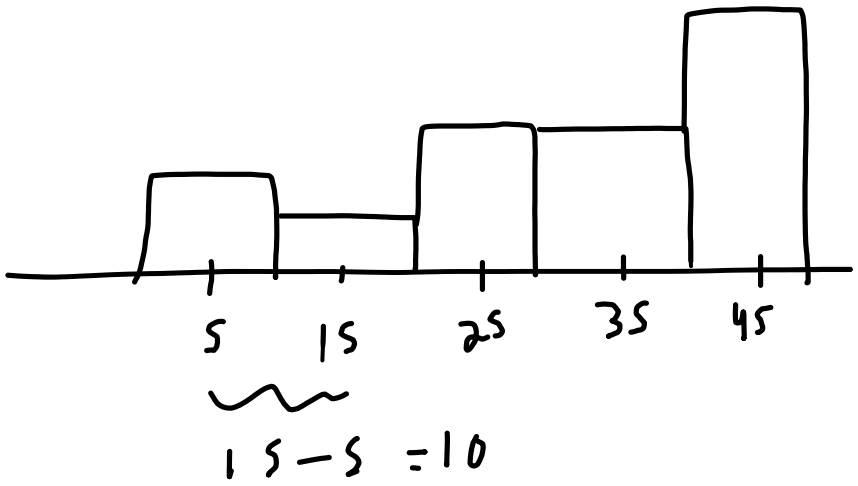
2) How many games had 30 or more points (score ≥ 30)? *4 + 7 = 11*

3) What percentage of games had less than 20 points (score < 20)? *0.1 + 0.15 = 0.25 or (3 + 2) / 20*

Histograms again

Histograms by Midpoints

- Will also see classes defined by the **Midpoints**, rather than the lower and upper bounds
  - Midpoint = (Lower limit + Upper limit) / 2
- This is how we will make histograms on the Lab in Excel
- Can find the class width by subtracting consecutive midpoints!

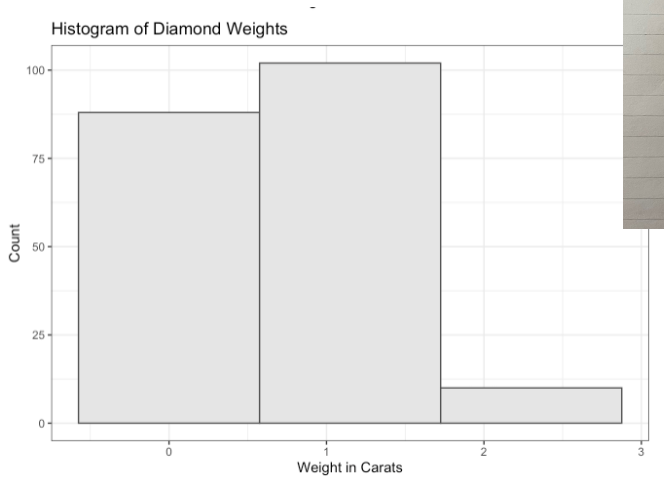


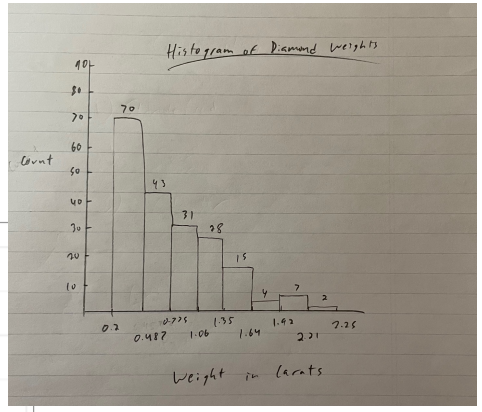
Bin	Midpoint	Frequency	Relative Frequency
0-10	*(0 + 10) / 2 = 5*	*3*	*0.15*
10-20	*(10+20) / 2 = 15*	*2*	*0.1*
20-30	*25*	*4*	*0.25*
30-40	*35*	*4*	*0.25*
40-50	*45*	*7*	*0.35*
Total:		*20*	*1*

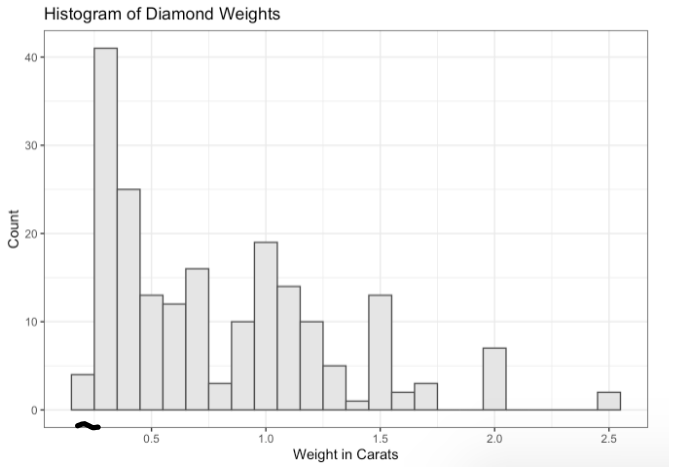
Histograms

<u>Choosing the Bins</u>

- Again bin selection has a big impact on the final graph.
- Don’t want too few bins, cause then you don’t really get any info…
- More bins is better (to an extent).
  - Shows the data more closely, can see more individual spikes.



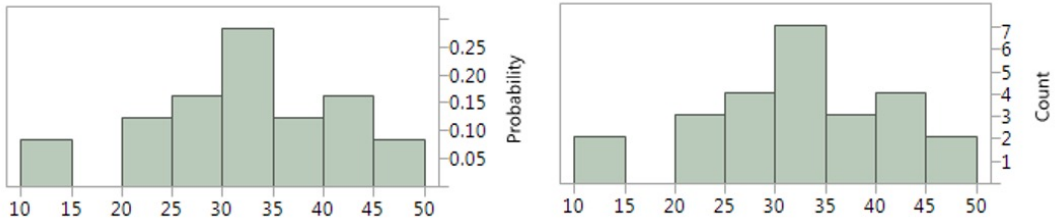




Relative Frequency Histogram

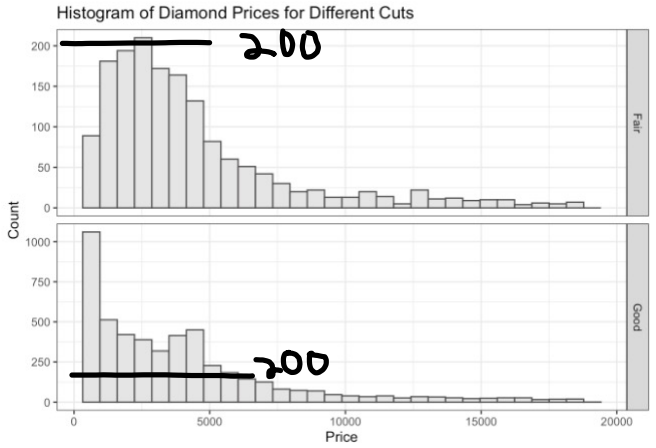
Relative Frequency Histogram

- Same as a regular histogram, except now the Relative Frequencies (proportions, probabilities, %) are on the y-axis.
- Sum of all the height equals 1.



Advantage

- Useful when comparing groups with DRASTICALLY DIFFERENT sample sizes.
  - Just using counts could be misleading….



Cut	Frequency
Fair	1610
Good	4906

- Using proportions puts the bars on the same scale, making it a fair (visual) comparison

Stem and Leaf Plots

Stem-and-leaf plot for the hotel rate data

5	0 4
5	5 8
6	0 1 1 3 4
6	5 8 9 9
7	0 1 2 2 2 3 3 3 3 4
7	5 5 5 5 5 6 7 7 7 7 8 8 8 8 9 9 9 9
8	0 1 1 1 3 4
8	5 5 6 9 9
9	3 3 4
9	0 1

Stem = tens
Leaf = ones

Stem and Leaf Plot

- (Weird) way to display **quantitative** data.
- Best for SMALL datasets.
- Two-sided chart that separates data values by digits.
- Displays the **shape** of the distribution and displays **all data values**.

Stem	Leaf	Count
4	88	2
4	2444	4
3	558	3
3	0002224	7
2	5699	4
2	022	3
1		
1	14	2

Data: 48, 48, 42, 44, 44, 44, 35, 35, 38, 30, 30, 30,…., 14

Stem (tens place)
Leaves (ones place)

Back-to-Back Stemplot

- Used to compare two distributions

Girls		Boys
9, 2	9	
6, 1, 0, 0	10	5, 8, 9
8, 7	11	0, 1, 1, 7
6, 6, 5, 5, 5, 4, 2	12	3, 7, 7, 8
7, 1, 0	13	3, 3, 4, 4, 6, 9
9, 8	14	4, 4, 5
8, 6, 2, 0, 0	15	0, 1, 2, 3, 7
7	16	2, 2, 2, 5, 8
	17	1, 6
8, 0	18	2, 8
9	19	5
4	20	

https://www.ck12.org/statistics/two-sided-stem-and-leaf-plots/

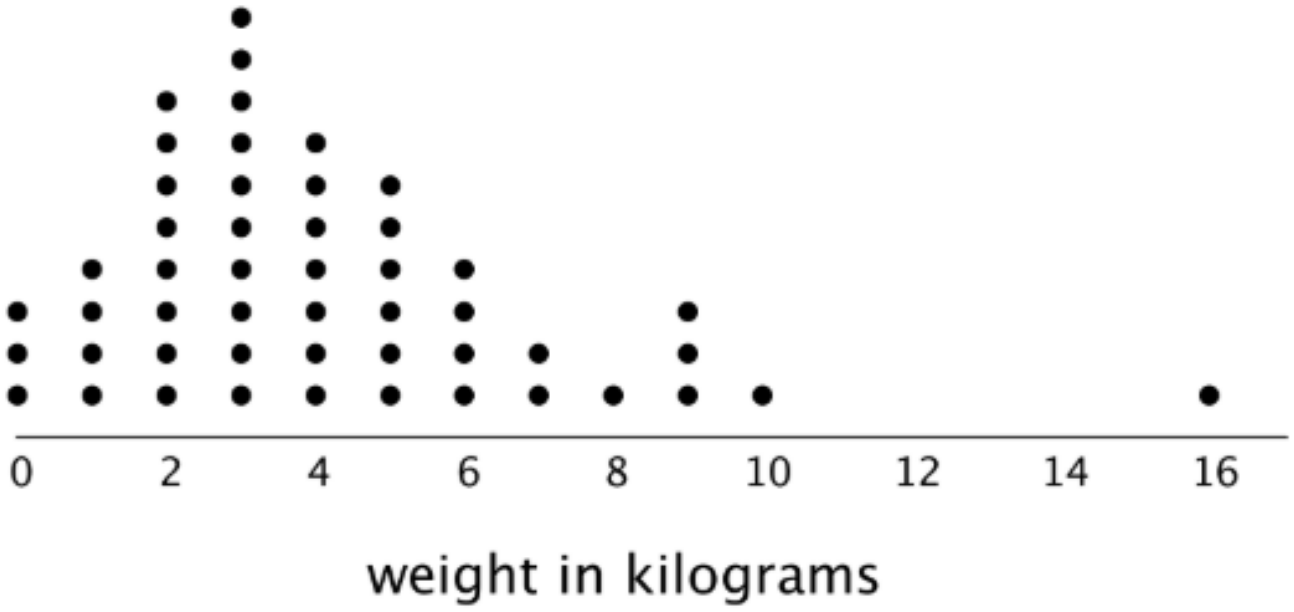
Dot Plots

Dot Plot

- Quick and easy method for organizing quantitative data, dots on a number line.
- Best for SMALL datasets.
- Displays the **shape** of the distribution and displays **all data values**.
- Can see “outliers”.

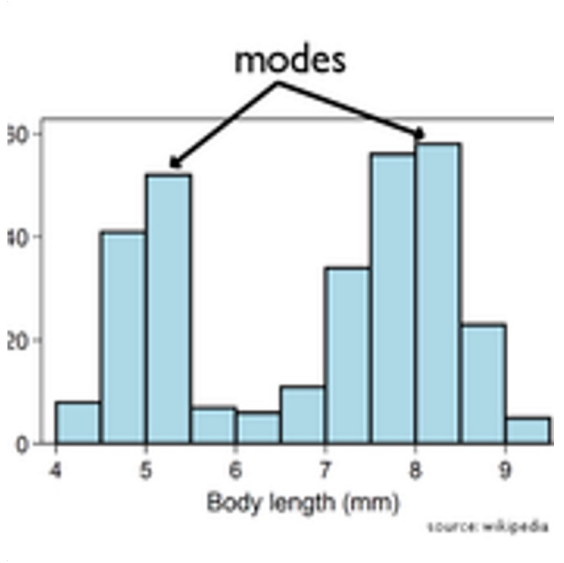
How to Construct

- Just stack dots above the value…

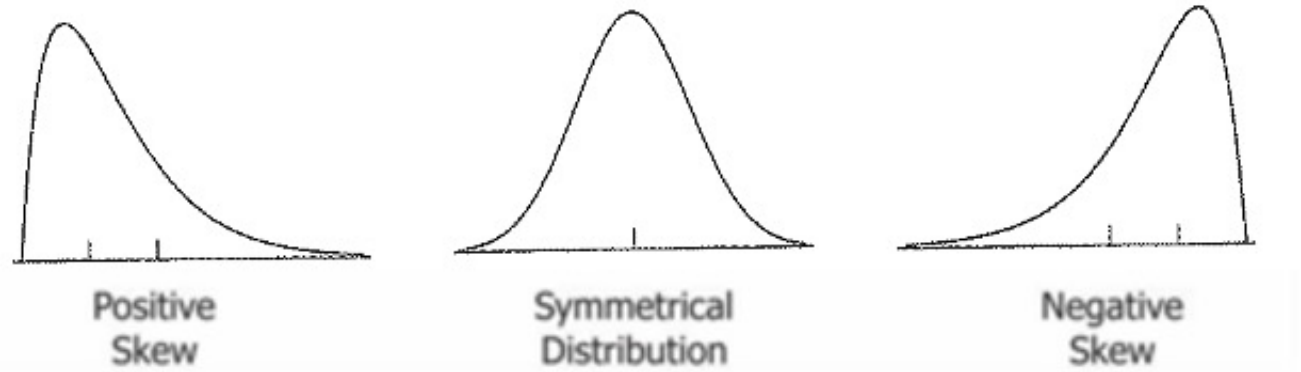


https://math.libretexts.org/Bookshelves/Arithmetic_and_Basic_Math

Measures of Shape



- Define how the data is distributed.
- Up to 2 things to report.
    - SHAPE: Is the data symmetric, right (positive) skewed, or left (negative) skewed?
    - MODALITY: Is the data unimodal, bimodal or multimodal?
        - Depending on the type of graph, this won't always be known. If you can't tell, don't report.



<u>Another shape</u>

Uniform

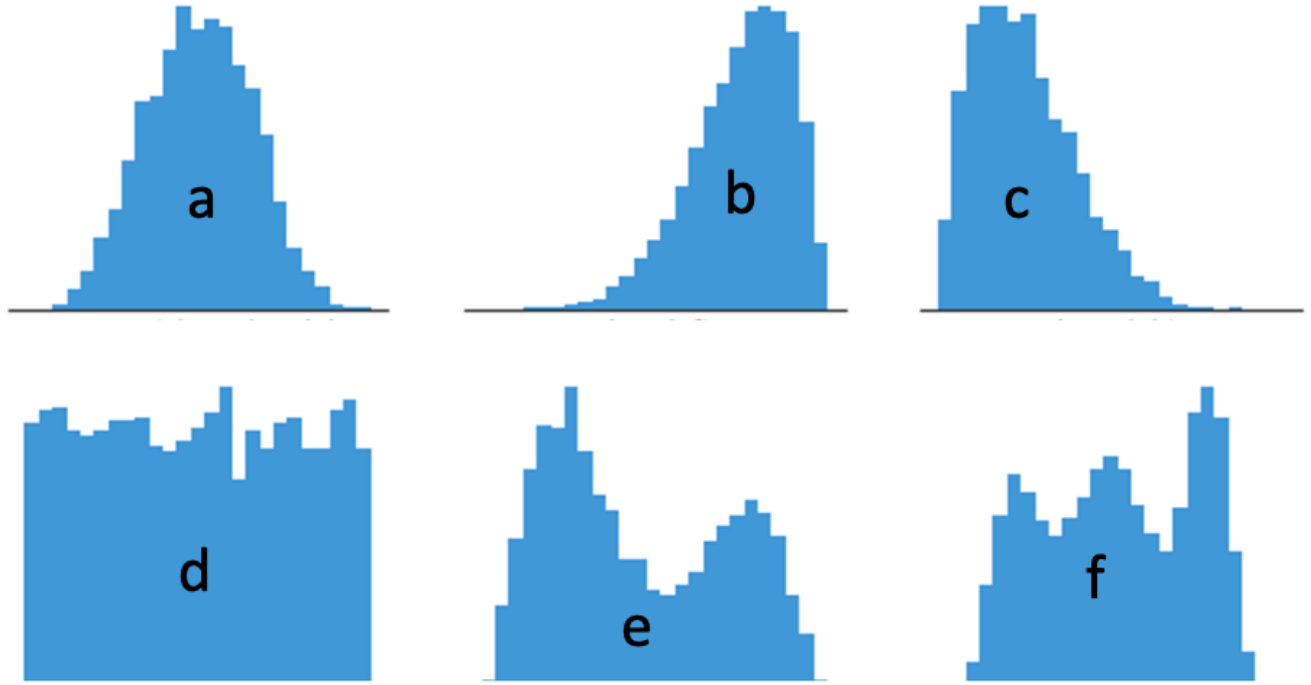
- Roughly **same** height across.

http://mathcenter.oxford.emory.edu/site/math117/shapeCenterAndSpread/

LCQ: Measures of Shape

Describe the SHAPE and MODALITY of each histogram

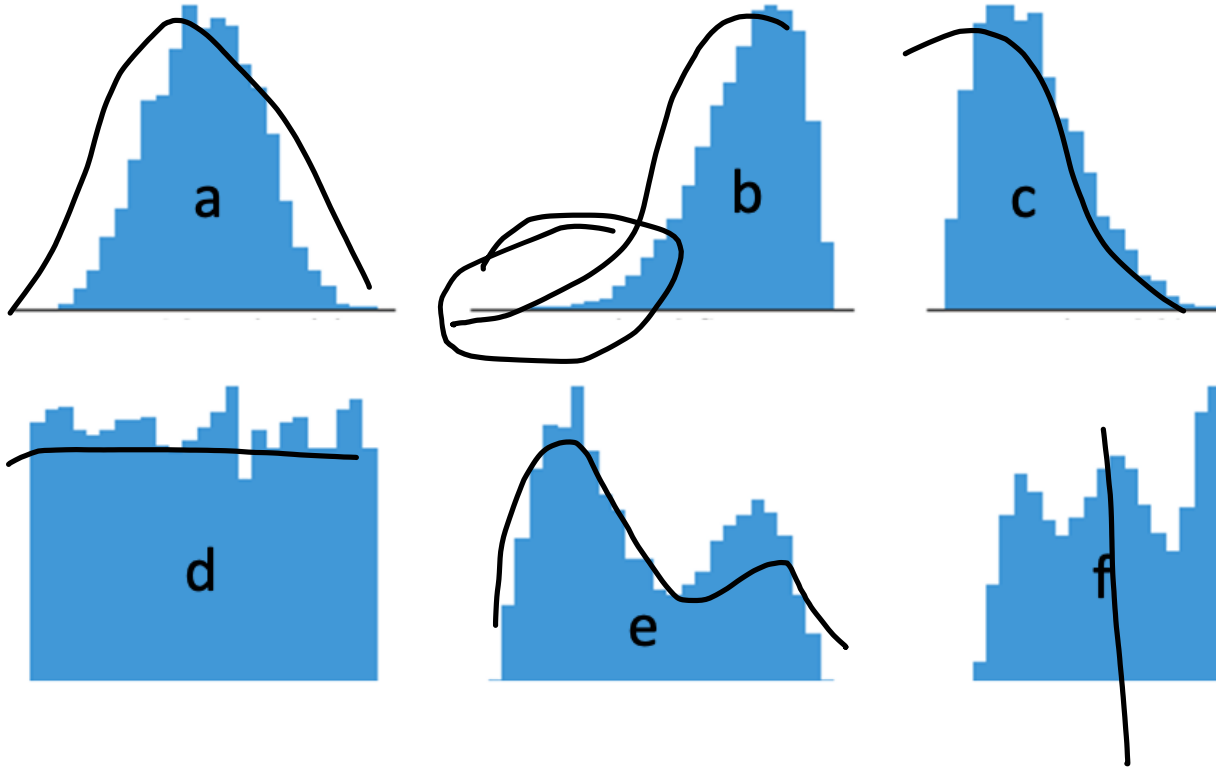
a)
b)
c)
d)
e)
f)



https://chartio.com/learn/charts/histogram-complete-guide/

LCQ: Measures of Shape

Describe the SHAPE and MODALITY of each histogram



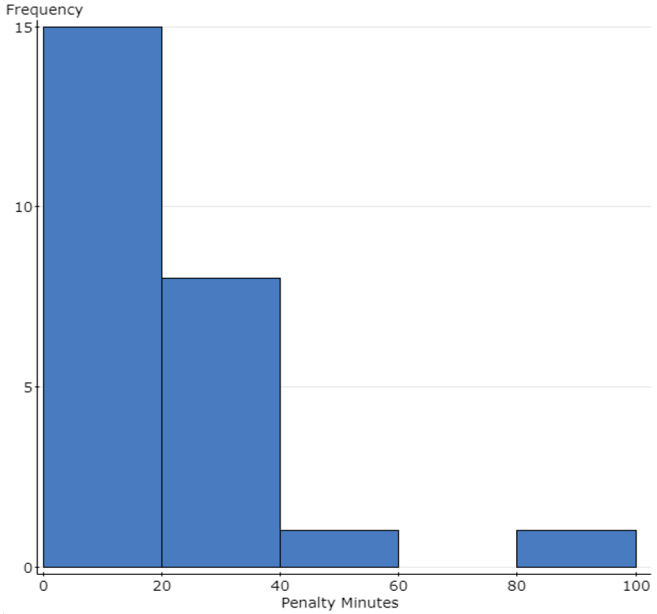
a) Symmetric and unimodal
b) Negative (left) skew and unimodal
c) Right skewed (positive) and unimodal
d) Roughly uniform (implies symmetric); can't say anything about the modality
e) Not symmetric and bimodal
f) Non symmetric (maybe roughly symmetric) and multimodal

https://chartio.com/learn/charts/histogram-complete-guide/

Overall Example

- Setup: Penalty minutes per player during the 14-15 season for the Miami University RedHawks Ice Hockey Team are located in the table below as well as in the dataset “2014 RedHawk Hockey”.

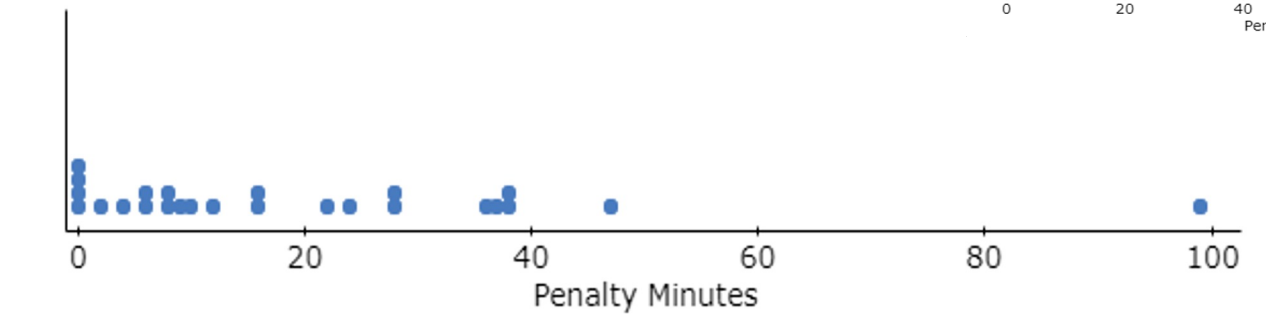
- Here are 3 different graphs to display the same info. *Notice what information you can get (and what you can’t) from each display.*



Variable: Penalty Minutes

Decimal point is 1 digit(s) to the right of the colon.
Leaf unit = 1

```
0 : 000024
0 : 66889
1 : 02
1 : 66
2 : 24
2 : 88
3 :
3 : 6788
4 :
4 : 7
5 :
5 :
6 :
6 :
7 :
7 :
8 :
8 :
9 :
9 : 9
```



- *All display shape (right skewed) and modality (unimodal)*
- *Histogram only shows mode as 0 – 20, whereas stemplot and dotplot show its really just 0 – 5 or 0 (most exact)*
- *All show outliers, but with histogram we only know the range of the outlier (80 – 100), whereas stemplot and ~ dot plot shows us it is 99*
- *Stemplot and ~ dotplot show all values*

Time Series Plots

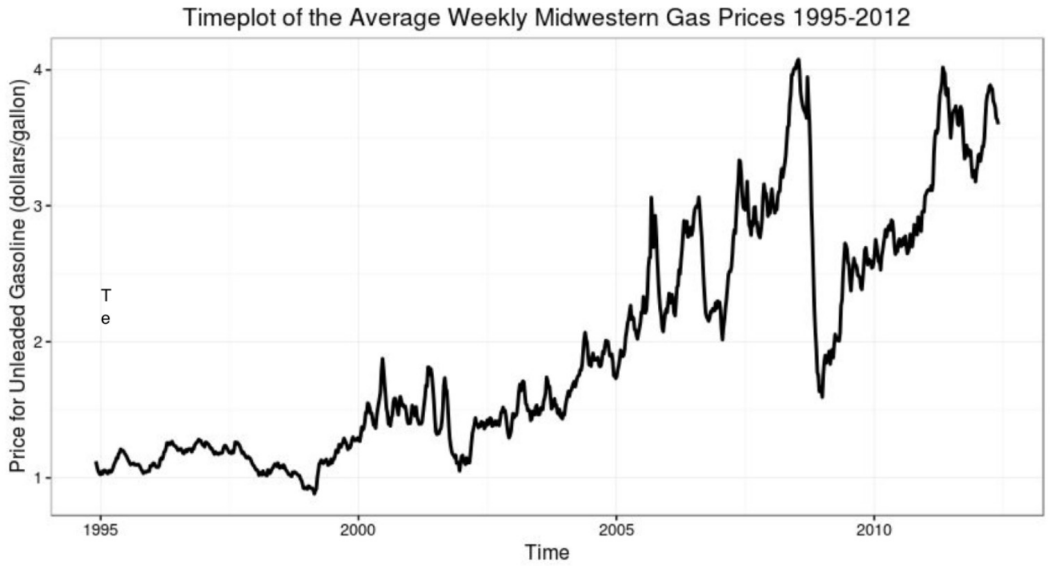
Time Series Plot

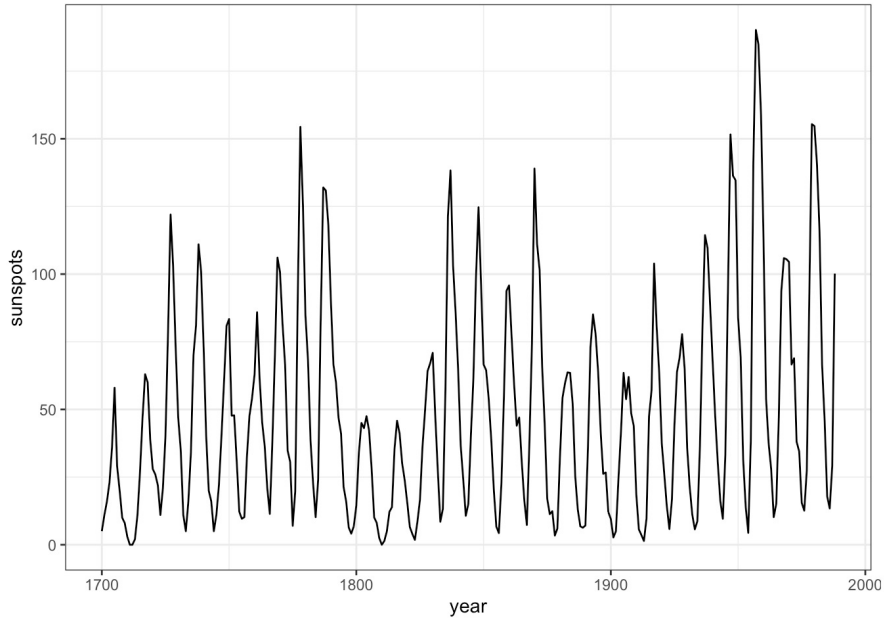
- Displays changes in a quantitative variable over time (aka **time series data**)**.**
- Time values on x-axis and values on y-axis.
  - Time is measured over equally spaced increments, e.g. days, months, years, etc.

- Best way to see trends (long-term upwards or downwards) over time!
- Also shows seasonal variation (cyclical pattern)!
  - This can be interpreted as change over time that has a regular pattern that repeats.
  - Examples: Hourly temperatures, monthly gym enrollment

How to Construct

- Line graph (connect the dots) with time values on x-axis and values on y-axis.



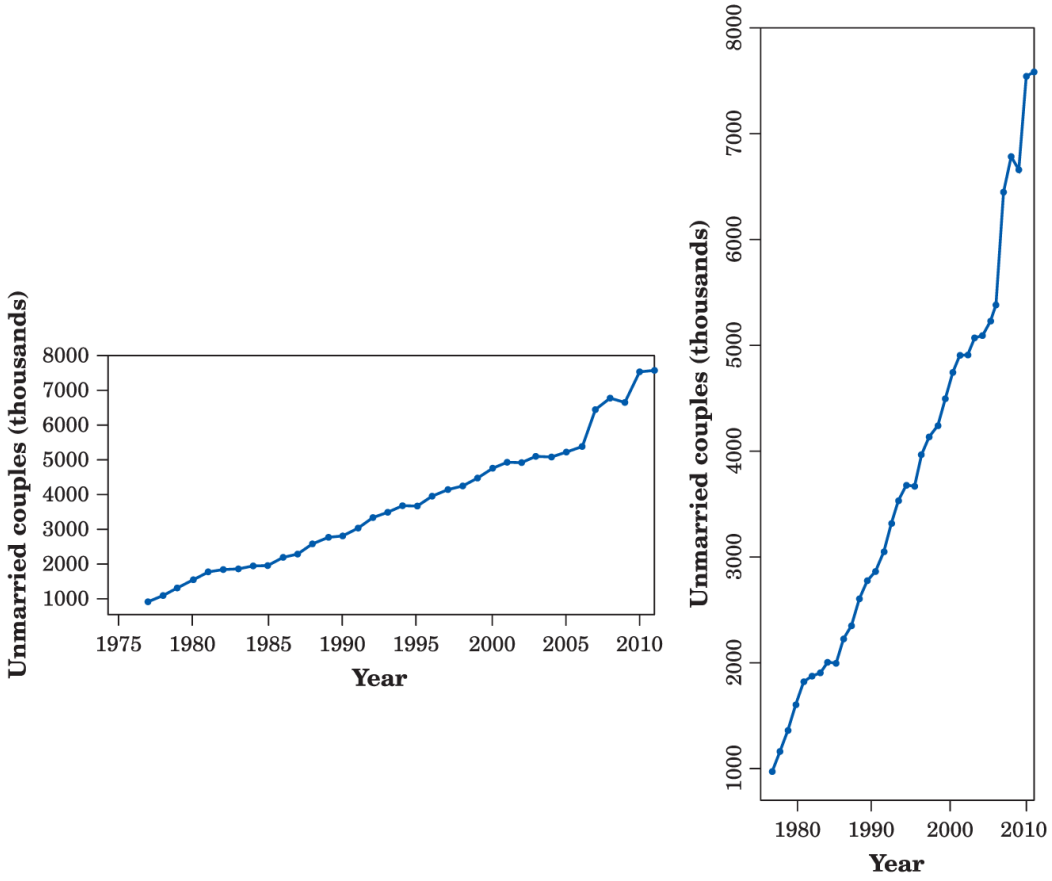


https://wwwbis.sidc.be/silso/datafiles

Scales of Time Series Plots

Scales of Time Series Plots

- This is something that you need to pay attention to when interpreting a time series plot.
- Very easy to manipulate axes to change the interpretation of a visual.
  - Horizontal stretches make the changes over time seem more gradual.
  - Whereas vertical stretches make make changes seem way more drastic.



Moore/Notz, *Statistics: Concepts and Controversies*, 10e, © 2020 W. H. Freeman and Company

PROBLEM SESSION!!!!!!!!!!

Problem #1

As part of the marketing team at an Internet music site, you want to understand who your customers are. You send out a survey to 25 customers asking for demographic information. One of the variables is the customer's age.

20	38	35	30	22	34	44	44	29	35	30	26	48
30	25	32	22	42	14	32	29	48	44	11	32	

a) Make a histogram of the data using a bar width of 10 years.

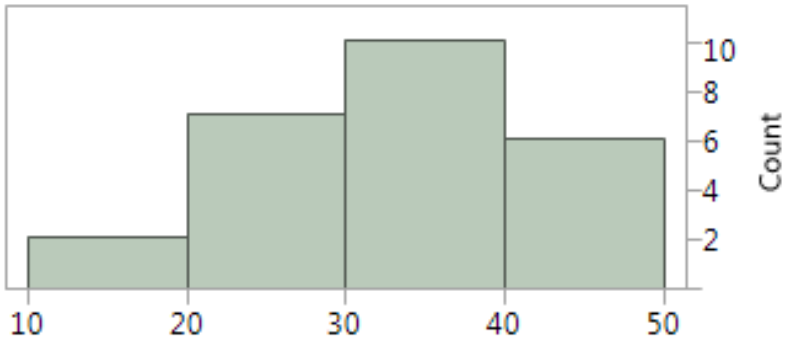
b) Make a histogram of the data using a bar width of 5 years.

c) Make a relative frequency histogram of the data using a bar width of 5 years.

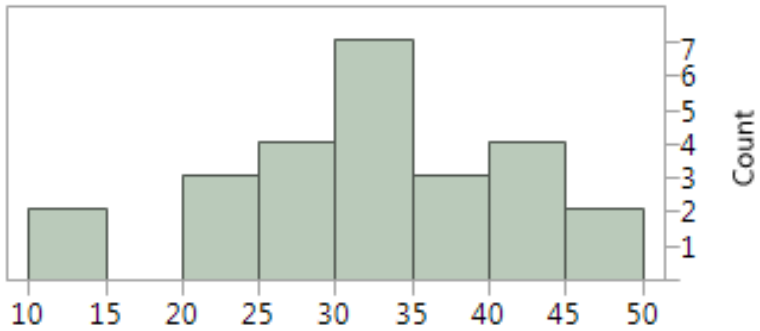
d) Make a stem-and-leaf plot of the data.

Problem #1 Solution

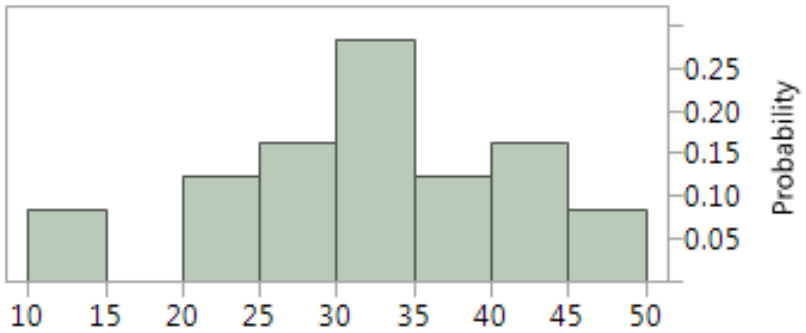
a)



b)



c)



d)

Stem	Leaf	Count
4	88	2
4	2444	4
3	558	3
3	0002224	7
2	5699	4
2	022	3
1		
1	14	2

1|1 represents 11

Problem #3

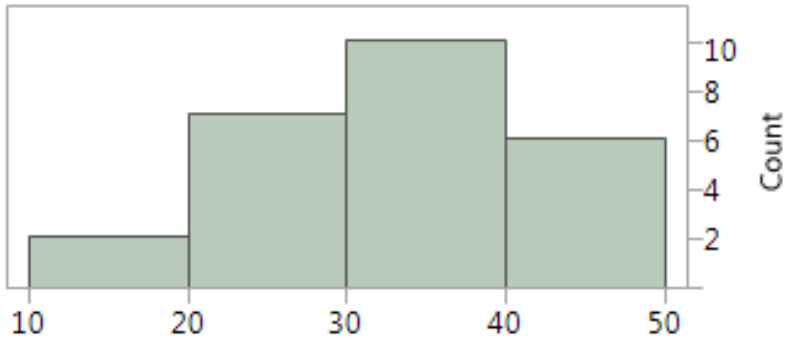
For the histogram that you made in Exercise 1a:

a) Is the distribution unimodal or multimodal?

b) Where is (are) the mode(s)?

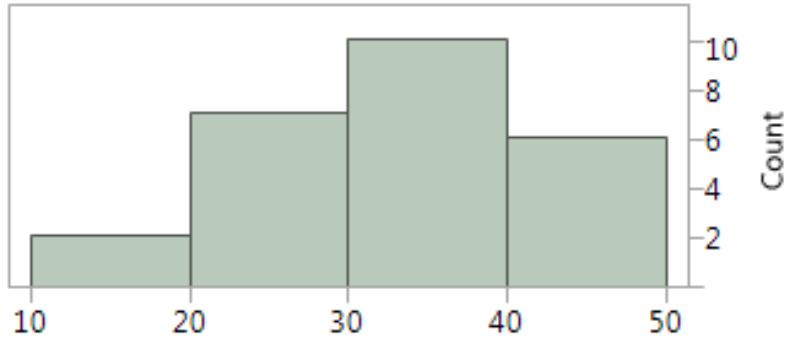
c) Is the distribution symmetric?

d) Are there any outliers?



Problem #3 Solution

a) Unimodal
b) Around 35
c) Fairly symmetric
d) No outliers



Problem #78

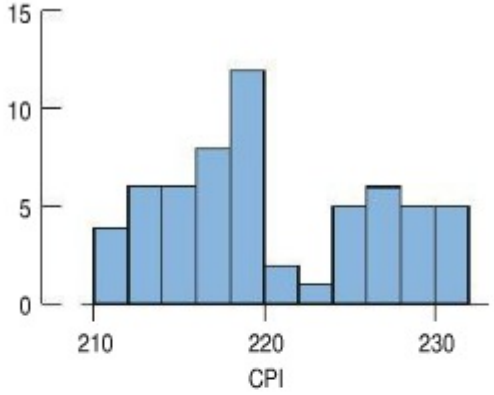
Here is a histogram and time series plot of the monthly CPI as reported by the Bureau of Labor Statistics from 2008 through 2013.

a) What features of the data can you see in the histogram that aren't clear from the time series plot?

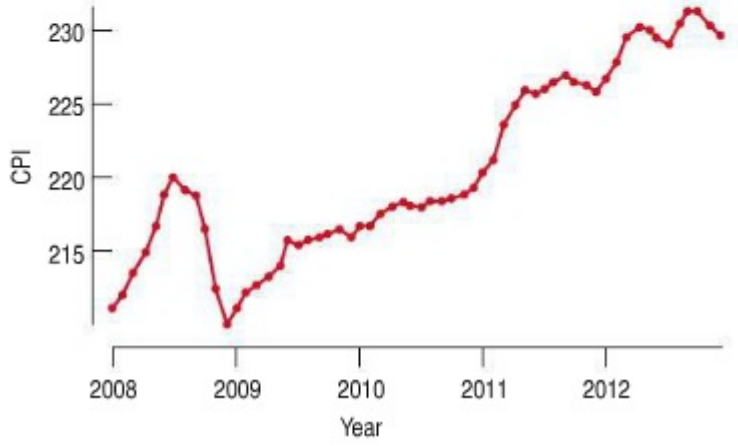
b) What features of the data can you see in the time series plot that aren't clear in the histogram?

c) Which graphical display seems the more appropriate for these data? Explain.

d) Write a brief description of monthly CPI over this time period.



Here is the time series plot for the same data.



Problem #78 Solution

a) The frequency of the different CPI values
b) The trend of the values over time for the CPI
c) The time series plot
d) The monthly CPI increased until July 2008 and decreased slightly for August and September. Then decreased sharply during the months of October through December where it hit its all-time low. The CPI had a slow, but steady increase until about December 2010 when it started increasing sharply. CPI have increased overall, except for slight dips from Oct-Dec 2011 and March – July 2012.