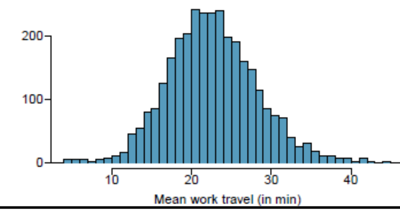
ECO239 Statistics I

Week 3

Graphs for Numerical Data
Histogram

- Bar Graph with Frequencies of each interval (No gap between bars)
- Intervals in x-axis, frequencies in y-axis.



Create a Histogram

- **First step**: Create frequency distribution table with k intervals with width w.

(you decide k, then calculate w.
See last week's slides for details)

- **Second step**: Create a bar chart (Histogram) without any gaps between bars.

- Height

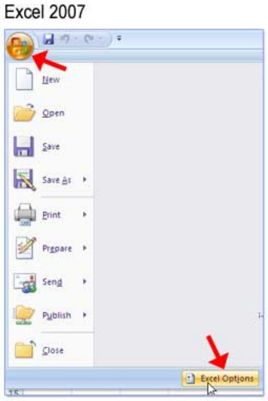
Histogram in Excel

- First Data Analysis tool should be installed to your excel if you cannot find "data analysis" tab on "data" tab.

- To install data analysis tool, press "start button" on excel -> Excel options->Add Ins -> Analysis Took Pak (select) -> press "Go". Data Analysis tool should appear on Data tab. (for Excel 2007).

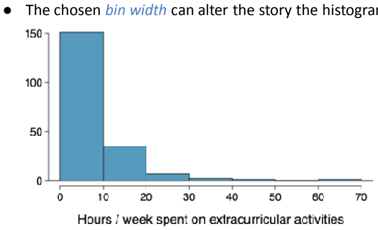
For other version of excel, google and learn "how to add "Data Analysis"tool on Excel ____".

Excel 2007



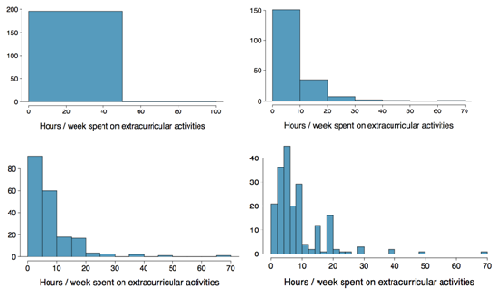
Histograms

- Histograms provide a view of the *data density*. Higher bars represent where the data are relatively more common.
- Histograms are especially convenient for describing the *shape* of the data distribution.
- The chosen *bin width* can alter the story the histogram is telling.



Bin Width

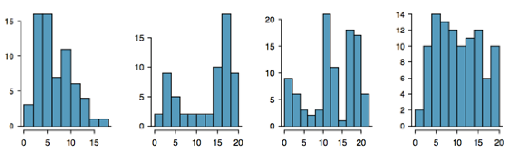
Which one(s) of these histograms are useful? Which reveal too much about the data? Which hide too much?



- Too Few Intervals (Wide Class Intervals)
  – May compress variation too much
  – Yield a blocky distribution
  – May observe important patterns of variation
- Too Many Intervals (Narrow Class Intervals)
  – May have many empty classes
  – Could give a poor indication of how frequency varies across classes.

Shape of a Distribution: Modality

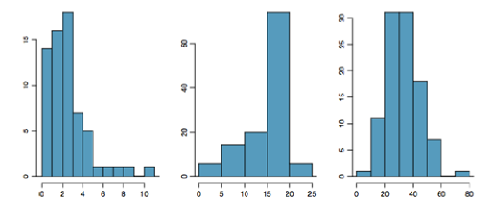
Does the histogram have a single prominent peak (*unimodal*), several prominent peaks (*bimodal/multimodal*), or no apparent peaks (*uniform*)?



Note: In order to determine modality, step back and imagine a smooth curve over the histogram -- imagine that the bars are wooden blocks and you drop a limp spaghetti over them, the shape the spaghetti would take could be viewed as a smooth curve.

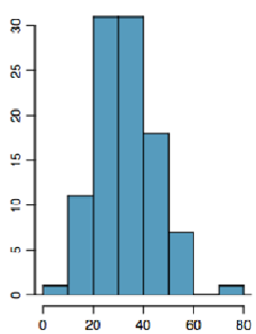
Shape of a Distribution: Skewness

Is the histogram *right skewed*, *left skewed*, or *symmetric*?

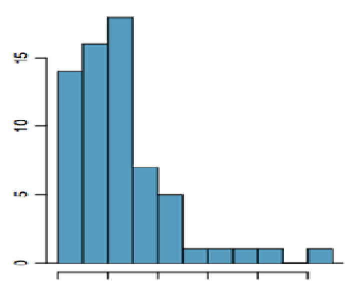


Note: Histograms are said to be skewed to the side of the long tail.

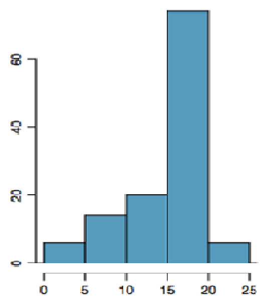
Symmetric Distribution



Positively Skewed (Skewed to the right)

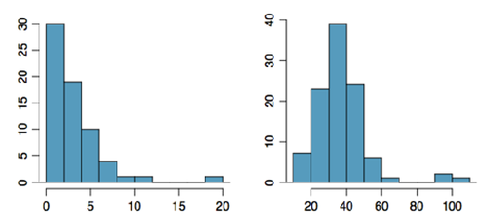


Negatively Skewed (Skewed to the left)



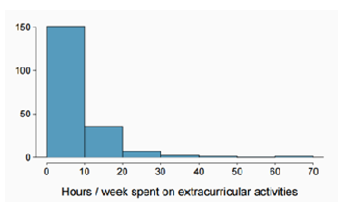
Shape of a Distribution: Unusual Observations

Are there any unusual observations or potential *outliers*?



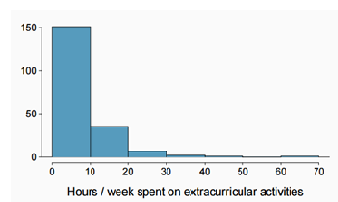
Extracurricular activities

How would you describe the shape of the distribution of hours per week students spend on extracurricular activities?



Extracurricular activities

How would you describe the shape of the distribution of hours per week students spend on extracurricular activities?



Unimodal and right skewed, with a potentially unusual observation at 60 hours/week.

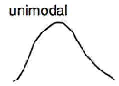
Commonly observed shapes of distributions

Modality

Commonly observed shapes of distributions

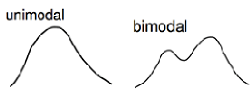
Modality

unimodal



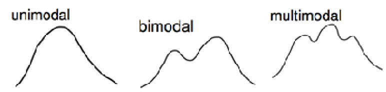
Commonly observed shapes of distributions

Modality



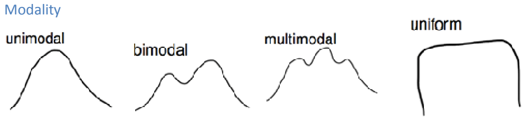
Commonly observed shapes of distributions

Modality



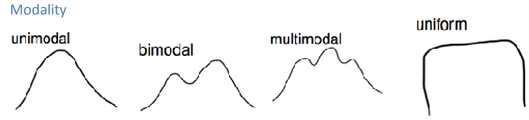
Commonly observed shapes of distributions

Modality



Commonly observed shapes of distributions

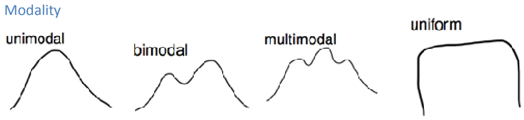
Modality



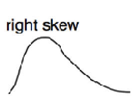
Skewness

Commonly observed shapes of distributions

Modality

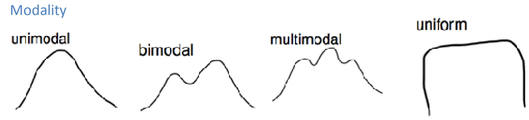


Skewness

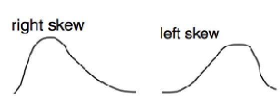


Commonly observed shapes of distributions

Modality

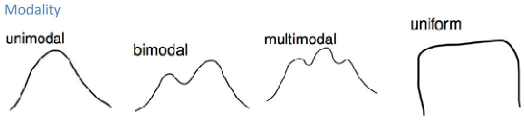


Skewness

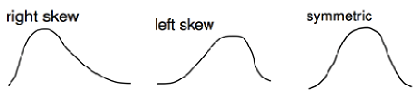


Commonly observed shapes of distributions

Modality



Skewness



Practice

Which of these variables do you expect to be uniformly distributed?

(a) weights of adult females
(b) salaries of a random sample of people from Ankara
(c) house prices
(d) birthdays of classmates (day of the month)

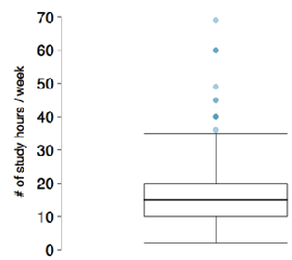
Practice

Which of these variables do you expect to be uniformly distributed?

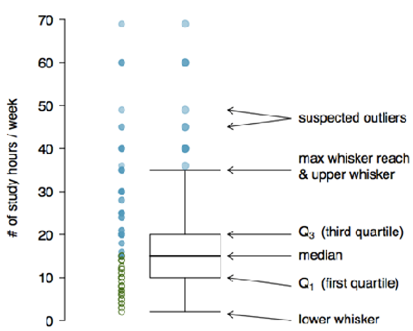
(a) weights of adult females
(b) salaries of a random sample of people from Ankara
(c) house prices
(d) birthdays of classmates (day of the month)

Box Plot

The box in a *box plot* represents the middle 50% of the data, and the thick line in the box is the median.

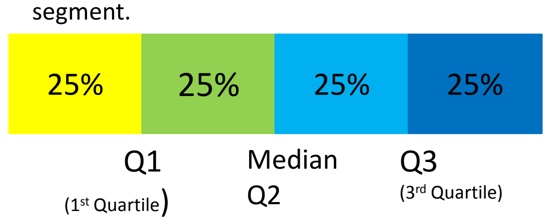


Anatomy of a Box Plot



Interquartile Range (to create Box plot)

- Quartiles: split the ranked data into 4 segments with an equal number of values per segment.



Interquartile Range = Q3 – Q1

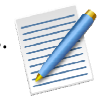
Q1 locates in ¼ (n+1) position
(25% below, 75% above)

Q2 locates in ½(n+1) position
(50% below, 50% above)

Q3 locates in ¾(n+1) position
(75% below, 25%above)

IQR: Practice 1

- Data { 0, 2, 3, 7, 9, 11, 13, 14, 16, 17, 21}
- n = 11
- Derive Q1, Q2 and Q3 values.



Data { 0, 2, 3, 7, 9, 11, 13, 14, 16, 17, 21}

- Q1 location = ¼(11+1) = 3 (=> 3^{rd} value = 3)
- Q2 location = ½(11+1) = 6 (=> 6^{th} value = 11)
- Q3 location = ¾(11+1) = 9 (=> 9^{th} value = 16)

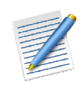
- Q1= 3
- Q2=11
- Q3=16

- IQR = 16-3 = 13

IQR: Practice 2

Data { 0, 2, 3, 7, 9, 11, 13, 14, 16, 17, 21, 22, 23}
n= 13

Derive IQR.

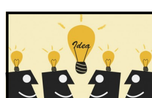


Data { 0, 2, 3, 7, 9, 11, 13, 14, 16, 17, 21, 22, 23}

- Q1 location = ¼ (13+1) = 14/4=7/2=3.5
- Q2 location = ½(13+1)=7
- Q3 location = ¾(13+1)= 3*14/4 = 21/2=10.5

⇒Q1 = 3 + (7-3)*0.5 = 5
⇒Q2 = 11
⇒Q3 = 17+ (21-17)*0.5 = 19

⇒IQR = 19-5 = 14



IQR: Practice 3

Data { 0, 2, 3, 7, 9, 11, 13, 14, 16, 17, 21, 22}
n= 12

Derive IQR.

Data { 0, 2, 3, 7, 9, 11, 13, 14, 16, 17, 21, 22}

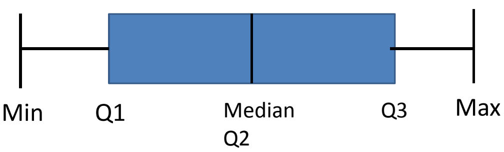
- Q1 location = ¼ (12+1) = 13/4=3.25
- Q2 location = ½(12+1)=13/2=6.5
- Q3 location = ¾(12+1)= (3*13)/4 = 39/4=9.75

⇒Q1 = 3 + (7-3)*0.25 = 4
⇒Q2 = 11+(13-11)*0.5=12
⇒Q3 = 16+ (17-16)*0.75 = 16.75

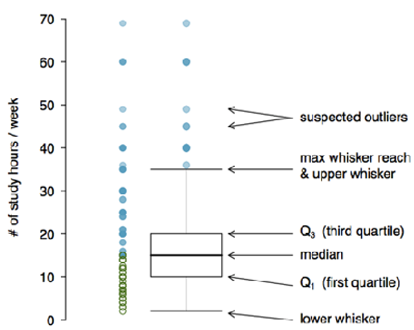
⇒IQR = 16.75-4 = 12.75

Five Number Summary

Min < Q1 < Median (Q2) < Q3 < Max

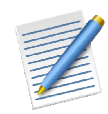


Now back to Box Plot



Now create a box plot

- Finalscore



NOTE: There is no “Box plot” function on Excel. Once we learn R program, we can draw one easily! For now, draw by hand.

Whiskers and Outliers

Whiskers of a box plot can extend up to

1.5 x IQR away from the quartiles.

- max upper whisker reach = Q3 + 1.5 x IQR
- max lower whisker reach = Q1 - 1.5 x IQR

Whiskers and Outliers

Whiskers of a box plot can extend up to 1.5 x IQR away from the quartiles.
max upper whisker reach = Q3 + 1.5 x IQR
max lower whisker reach = Q1 - 1.5 x IQR

IQR: 20 - 10 = 10

max upper whisker reach
= 20 + 1.5 x 10 = 35
max lower whisker reach
= 10 - 1.5 x 10 = -5

Whiskers and Outliers

Whiskers of a box plot can extend up to 1.5 x IQR away from the quartiles.

max upper whisker reach = Q3 + 1.5 x IQR
max lower whisker reach = Q1 - 1.5 x IQR

IQR: 20 - 10 = 10
max upper whisker reach = 20 + 1.5 x 10 = 35
max lower whisker reach = 10 - 1.5 x 10 = -5

A potential *outlier* is defined as an observation beyond the maximum reach of the whiskers. It is an observation that appears extreme relative to the rest of the data.

Outliers (cont.)

Why is it important to look for outliers?

Outliers (cont.)

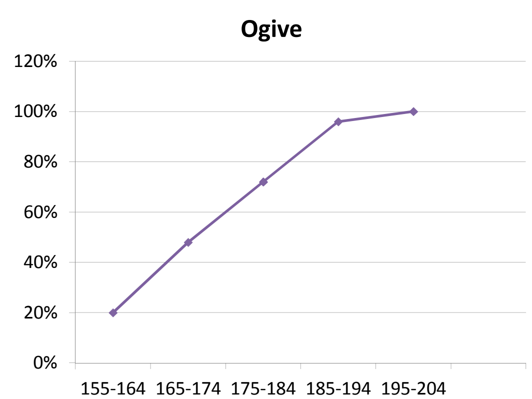
Why is it important to look for outliers?

- Identify extreme skew in the distribution.
- Identify data collection and entry errors.
- Provide insight into interesting features of the data.

Ogive: Cumulative Line Graph

- X-axis: interval
- Y-axis: relative cumulative frequency

height



Stem-and-Leaf Display

- A simple way to see the distribution details in a data set.

- Step1: sort data in ascending order
- Step2: observe the digits of the data
- Step3: separate the sorted data into
  - Leading digits (stem)
  - Trailing digits (leaves)

Practice

Data{ 31, 45, 48, 55, 57, 58, 67, 68, 73, 75, 78, 80, 82, 85, 88, 89, 91, 92, 95, 99}

Sorted data
2 digits
Leading digits = 10's digits
Trailing digits = 1's digits

Stem	Leaves (Leaf unit = 1)				
3	1				
4	5	8			
5	5	7	8		
6	7	8			
7	3	5	8		
8	0	2	5	8	9
9	1	2	5	9	

Case of 3 digits

Data { 223, 368, 378, 421, 468, 490, 526, 574, 647}

Sorted
3 digits
Round off the 2nd digits
223 => 220
368=> 370
Stem = 100's digit
Leave = 10's digit

Stem	Leave (Leaf unit = 10)		
2	2		
3	7	8	
4	2	7	9
5	3	7	
6	5		

Describing Relationship between Variables

1. Cross Table for Categorical Variables
2. Scatter Plot for Numerical Variables

Cross Table

- List number of observations for every combination of values for two categorical variables.
- R categories for 1st variables (rows)
- C categories for 2nd variables (columns)

Cross Table

	Investor A	Investor B	Investor C	Total
Stocks	46	55	27	128
Bonds	32	44	19	95
CD (certificate of deposit)	15	20	13	48
Savings	16	28	7	51
Total	109	147	66	322

Scatterplot

Scatterplots are useful for visualizing the relationship between two numerical variables.

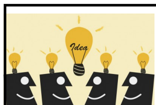
Example: Relationship between GPA and Final Exam Score

Data: GPA

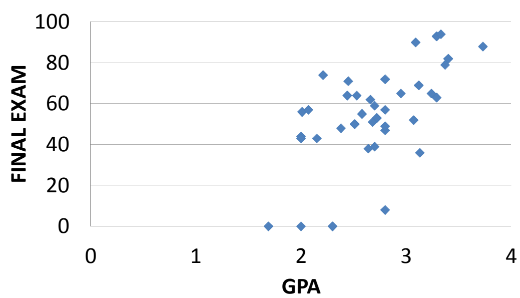
Q.What kind of relationship do you expect?

Q.How it can be plotted?

Q. How will it look like?



Scatter Plot



Quiz 1 (Oct.18)

Category	Frequency (sales last week)
iPhone	100
SONY	45
Samsong	75
HTC	20
TOTAL	240

Q:Generate a Parato Diagram