Descriptive Statistics cont.

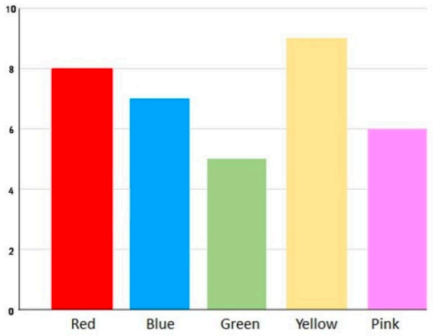
Measure of Relative Standing and Box Plots

Two Excel Files

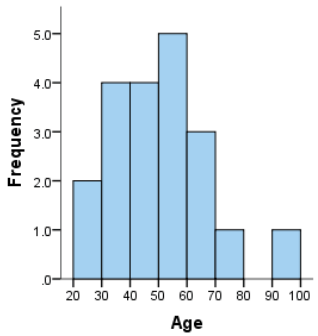
Instructor: Qasim Ali

Graphical Methods in Descriptive Statistics

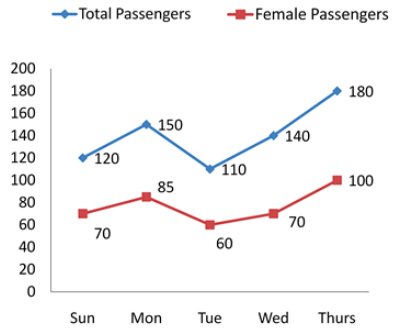
Bar Chart



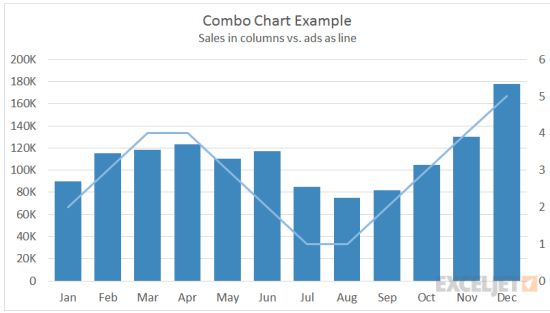
Histograms



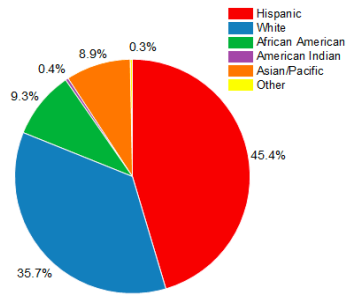
Line plot



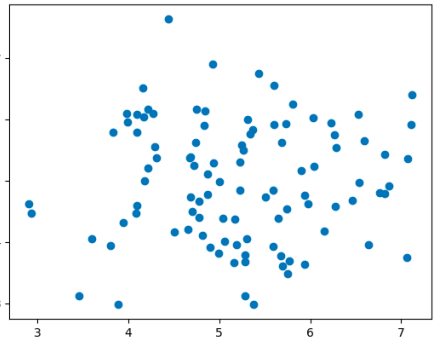
Combo Charts



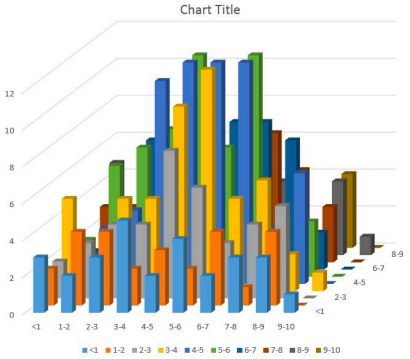
Pie Chart



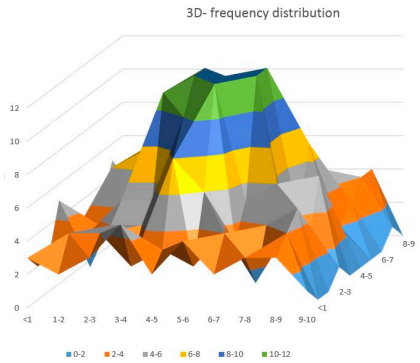
Scatter and bubble plot



3D Plot



Surface plot



Measures in Descriptive Statistics

Measure of Central Location

- We found the central location in the data
- ***Mean:*** Simple average
- ***Median:*** Middle value of the data
- ***Mode:*** Most frequently occurring value

μ	Population Mean
$\bar{x}$	Sample Mean
x_i	Values in the data
N	Total number of values
σ	Standard deviation

Measure of Variability

- We studied how to measure the spread of a dataset
- ***Range:*** Difference between the highest and lowest values
- **Standard Deviation:** measures the dispersion of a dataset relative to its mean

Roughly: "Average distance to the mean"

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2},$$

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}.$$

Measures of Relative Standing

Percentile P_k

- **P_k** percentile is a value which describes that ***k%*** of the data lies below this value.
- Remaining data that is above **P_k** is **100% – *k%***.

Quartiles

- Quartiles are percentiles that partition the data set into quarters.

Lower Quartile $Q_L = Q_1 = P_{25}$

Upper Quartile $Q_U = Q_3 = P_{75}$

Interquartile Range

- Difference between the Upper and Lower Quartile

$$IQR = Q_U - Q_L = Q_3 - Q_1 = P_{75} - P_{25}$$

- Let's take an example in excel worksheet.

Measures of Relative Standing

Locator of P_k

- Locator of $\boldsymbol{P_k}$ percentile in the data is:

$$\boldsymbol{L_k = \frac{(n+1)k}{100} = W + F}$$

where 'n' is the sample size, W is the integer part and F is the decimal part of L_k

Percentile Estimation

$$\boldsymbol{P_k = x_W + F(x_{W+1} - x_W)}$$

- If $F = 0$ then $L_k = W$ *and*
  $\boldsymbol{P_k} = \boldsymbol{x_W}$ = a value with the rank of $\boldsymbol{L_k}$

- If $F \neq 0$ then $L_k = W + F$ *and*
  $$\boldsymbol{P_k = x_W + F(x_{W+1} - x_W)}$$

Measures of Relative Standing

Example

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
6.8	7.4	7.9	8.2	**8.3**	**8.3**	8.4	8.8	9.1	9.8

We are required to find P_{25}?

To find P_{25}, we use the formula:

$$\boldsymbol{P_k = x_W + F(x_{W+1} - x_W)}$$

We require values of *F* and *W*.

$$L_{25} = \frac{\mathbf{(10+1)*25}}{\mathbf{100}} = \mathbf{2.75}$$

This implies *W* = 2 and *F* = 0.75. Finally,

$$\boldsymbol{P_{25} = Q_1} = x_2 + 0.75(x_3 - x_2) = 7.8$$

Find P_{50}

- $L_{50} = \frac{(10+1)50}{100} = \frac{550}{100} = 5.5$
- $$\begin{aligned}\boldsymbol{P_{50} = Q_2} &= x_5 + 0.5(x_6 - x_5)\\ &= 8.3 + 0.5(8.3 - 8.3)\\ &= 8.3 + 0.5\times 0\\ &= 8.3\end{aligned}$$

Find P_{75}

- $L_{75} = \frac{(10+1)75}{100} = \frac{825}{100} = 8.25$
- $$\begin{aligned}\boldsymbol{P_{75} = Q_3} &= x_8 + 0.25(x_9 - x_8)\\ &= 8.8 + 0.25(9.1 - 8.8)\\ &= 8.8 + 0.25\times 0.3\\ &= 8.875 \approx 8.9\end{aligned}$$

Interquartile range $= IQR = Q_U - Q_L = Q_3 - Q_1 = P_{75} - P_{25} = 8.9 - 7.8 = 1.1$

Measures of Relative Standing

Answer the following questions:

- Which *Measurement of Central Location (Mean, Median or Mode)* is equals to 50th Percentile or Q2?

  Answer: Median

- A percentile represents a value that is least amongst the *k*% of the data

  Answer: FALSE

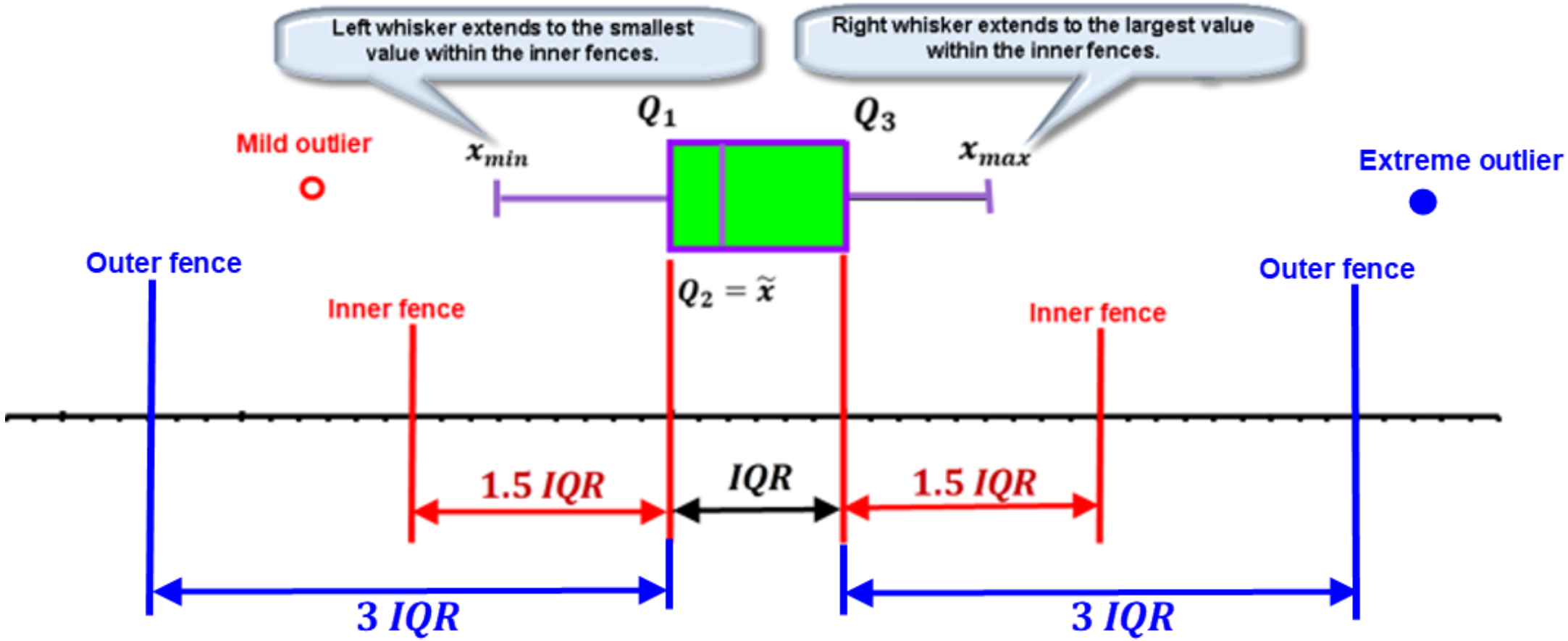
- In locator formula $L_k = W + F$, there are W values in a dataset that are less than the value of Percentile.

  Answer: TRUE

- Interquartile range is the difference between upper and lower quartile.

  Answer: TRUE

Box Plot


Left whisker extends to the smallest value within the inner fences.
Right whisker extends to the largest value within the inner fences.
Q_1
Q_3
Mild outlier
x_{min}
x_{max}
Extreme outlier
Outer fence
Inner fence
$Q_2 = \tilde{x}$
Inner fence
Outer fence
1.5 IQR
IQR
1.5 IQR
3 IQR
3 IQR

Box Plot

Example

Example

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
6.8	7.4	7.9	8.2	8.3	8.3	8.4	8.8	9.1	9.8

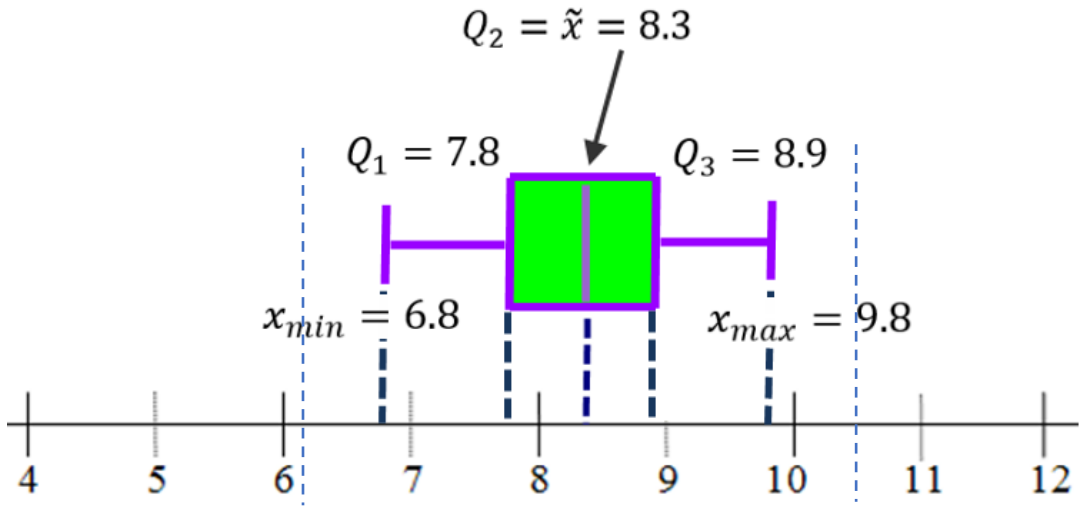
$x_{min} = 6.8, x_{max} = 9.8$

$Q_1 = 7.8, Q_2 = \tilde{x} = 8.3, Q_3 = 8.9$

1.5 Interquartile range $= 1.5\ IQR$

$= 1.5 \times (Q_3 - Q_1)$

$= 1.65$



Inner Fence:

Lower Limit $=$ LL $= Q_1 - 1.5IQR = 7.8 - 1.65 = 6.15$

Upper Limit $=$ UL $= Q_3 + 1.5IQR = 8.9 + 1.65 = 10.55$

Create a Box Plot in Excel

METHOD 1: EXCEL BOX PLOT OPTION (LINKED TO EXCEL FILE)

METHOD 2: STEP BY STEP APPROACH

To create your own box plot chart, the first step is to set up your data.

- Labels are not used in the chart. Let say the data is in column B and C with 13 values each.
- Enter the Box Plot Chart Formulas

Step 1: Calculate the quartile values

F4=MIN(B1:B13)	G4=MIN(C1:C13)
F5=QUARTILE(B1:B13,1)	G5=QUARTILE(C1:C13,1)
F6=MEDIAN(B1:B13)	G6=MEDIAN(C1:C13)
F7=QUARTILE(B1:B13,3)	G7=QUARTILE(C1:C13,3)
F8=MAX(B1:B13)	G8=MAX(C1:C13)

Step 2: Calculate quartile differences

=F5	=G5
=F6-F5	=G6-G5
=F7-F6	=G7-G6
=F8-F7	=G8-G7
=F5-F4	=G5-G4

Step 3: Create the Box Plot Chart

- Create a stacked column chart
- Convert the stacked column chart to the box plot style
- Hide the bottom data series
- Create whiskers for the box plot
- Color the middle areas

https://www.contextures.com/excelboxplotchart.html

https://support.microsoft.com/en-us/office/create-a-box-plot-10204530-8cdf-40fe-a711-2eb9785e510f

Box Plot

Answer the following questions:

- What is the basic use of box plot?

Answer: Skewness of Data