$\boxed{3.3}$ Position of a data point within the Group

> We seek to define the location of any given data point with in that data Point's Group

This tells us how to compare different data points

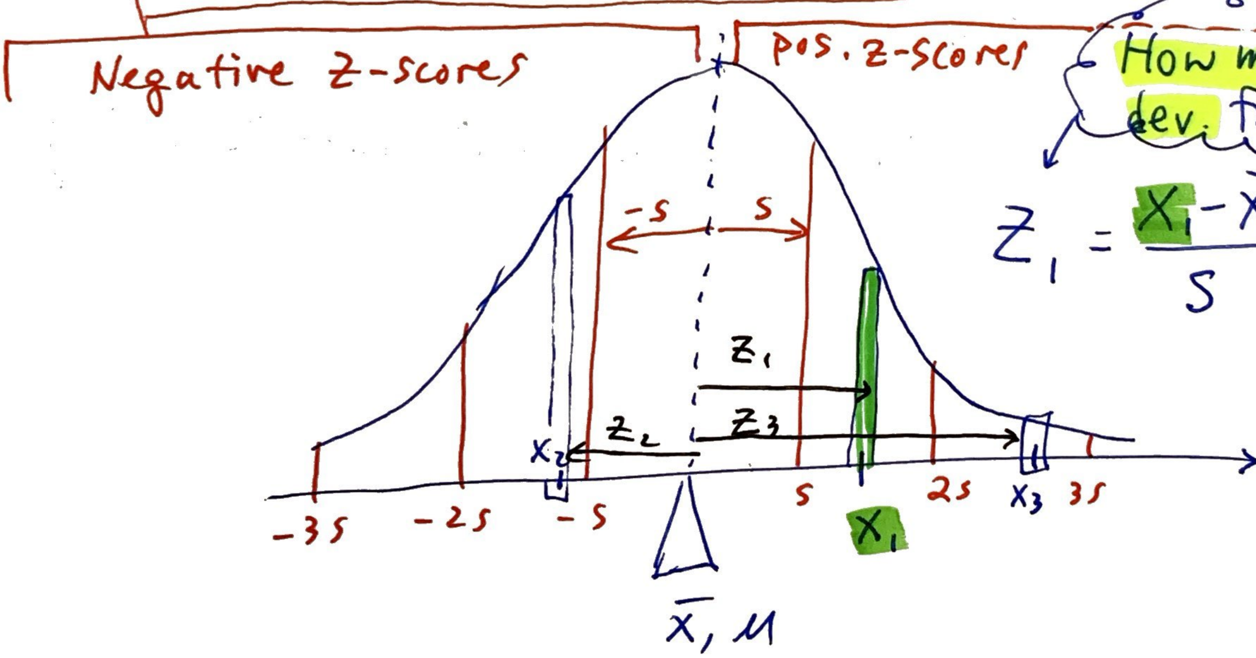
> $\boxed{Ex}$ Amongst their gender who is taller?
> a man @ 73" or a woman @ 68"

- To answer this question we introduce the

"Z - score"

$$\boxed{Z \equiv \frac{X - \mu}{\sigma}} \quad \text{and} \quad \boxed{Z = \frac{X - \bar{X}}{s}}$$

population census — sample from a population



How many std. dev. from the mean is this data point

$$Z_1 = \frac{X_1 - \bar{X}}{S}$$

Ex Statisticians found that in a pop of college men the average height is 69.4 inches with a std. deviation of 3.1 inches

For college women the mean is 63.8 in and a std. dev. of 2.8 inches

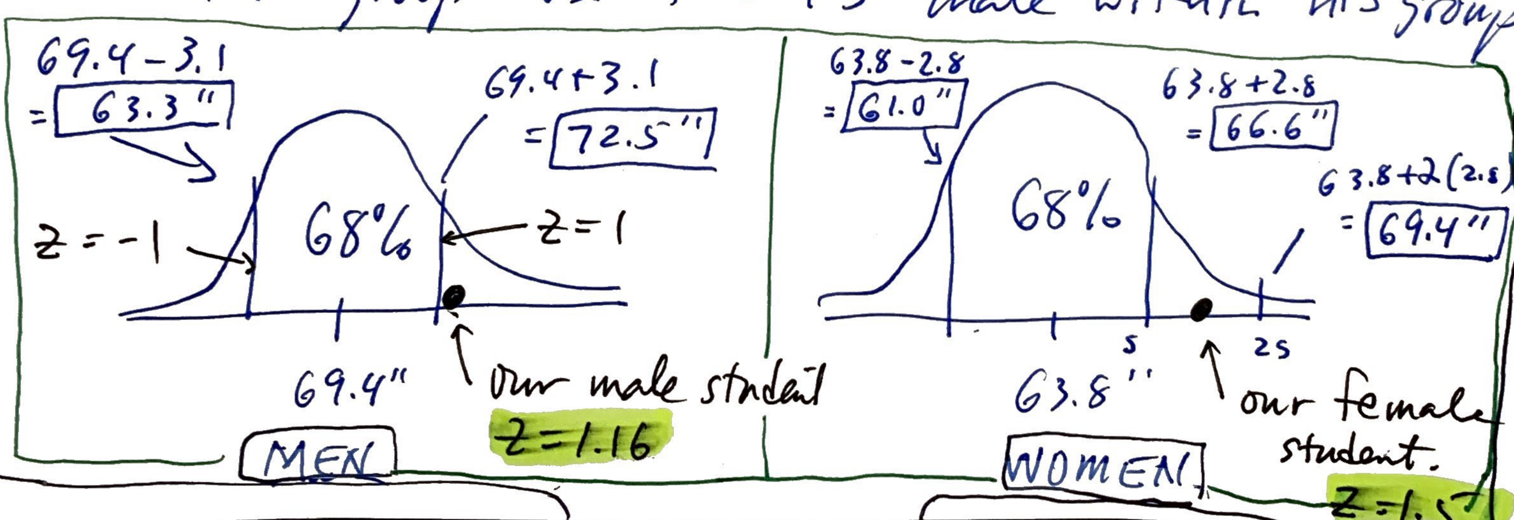
Q: who is taller with in their gender group: a 73" male or a 68" female?

- We can assume these groups has a unimodal and symmetric distribution thus allowing us to use z-scores $z = \frac{x-\mu}{\sigma}$

- The 73" male: $Z = \frac{73-69.4}{3.1} = \underline{1.16}$ std. dev from the mean

- The 68" female: $Z = \frac{68-63.8}{2.8} = \underline{1.50}$

Conclusion The 68" woman is taller within her group vs. the 73" male within his group.



Ex (cont.)

Q: Is a 75" man unusual?

i.e. is the z score of this man larger than 2?

$$z = \frac{75 - 69.4}{3.1} = 1.8$$ Not Unusual

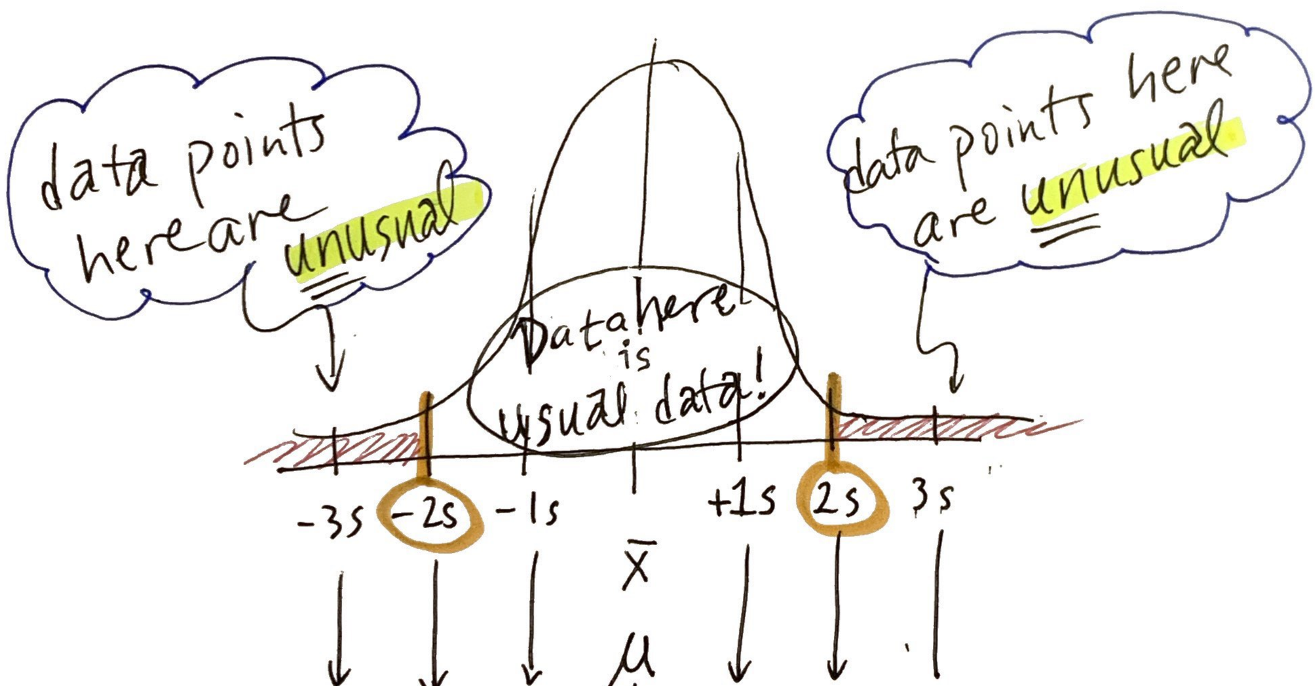
Q: Is a 5'4" woman unusual?

$$z = \frac{54 - 63.8}{2.8} = -3.5$$ very unusual

Unusual

Recall unusual is $\pm 2s$ or beyond;

this means that z-scores beyond ± 2 are unusual



Z-Score: -3 -2 -1 0 1 2 3

three S. Dev. above the mean

three Std. Dev. away below the mean

* Quartiles

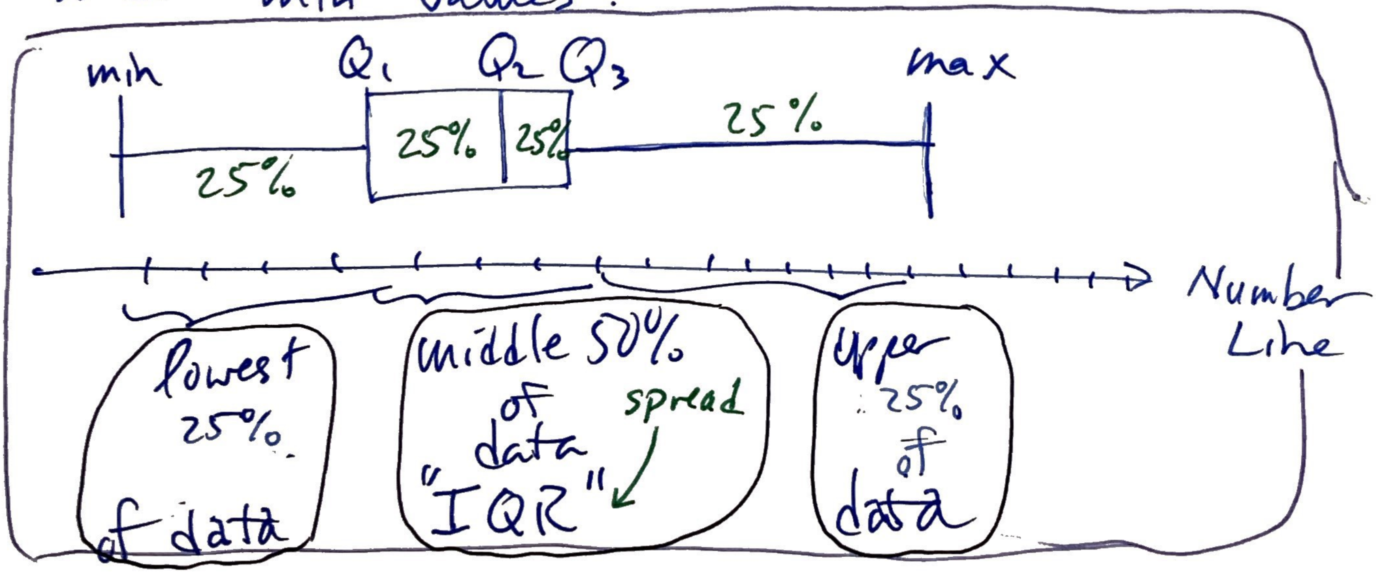
Def The **1st Quartile**, Q_1, is a data point that separates the lowest 25% of the data from the upper 75% of the Data.

The **2nd Quartile**, Q_2, is a data point that separates the lowest 50% of data from the upper 50% of data. (aka median)

The **3rd Quartile**, Q_3, is that data point that separates the lower 75% from the upper 25% of the data

* Box-Plot

A rectangle whoes edges are at Q_1 & Q_3 and that has "whiskers" that extend from these edges to the max and min values.



Def The IQR is called the Inter Quartile Range

$$\boxed{IQR = Q_3 - Q_1}$$

contains 50% of the middle data

* **5-number Summary**

min	______
Q_1	______
median	______
Q_3	______
max	______

The IQR is a measure of spread. It is effective for not only unimodal and symmeteric data distributions but also unimodal and skewed data.

EX Build a box plot for the follow data

- **Raw:** 65, 67, 71, 57, 51, 49, 44, 41, 59, 49, 42, 56, 45, 77, 44, 42, 45, 46, 100, 59, 53, 51

$N = 22$

- **Order:**

min → 41, 42, 42, 44, 44, 45 ← Q_1, 45, 46, 49, 49, $\boxed{51}$, $\boxed{51}$, 53, 56, 57, 59, 59 ← Q_3, 65, 67, 71, 77, 100 ← max

$\underbrace{41, 42, 42, 44, 44}_{5}$, 45, $\underbrace{45, 46, 49, 49, 51}_{5}$, $\underbrace{51, 53, 56, 57, 59}_{5}$, 59, $\underbrace{65, 67, 71, 77, 100}_{5}$

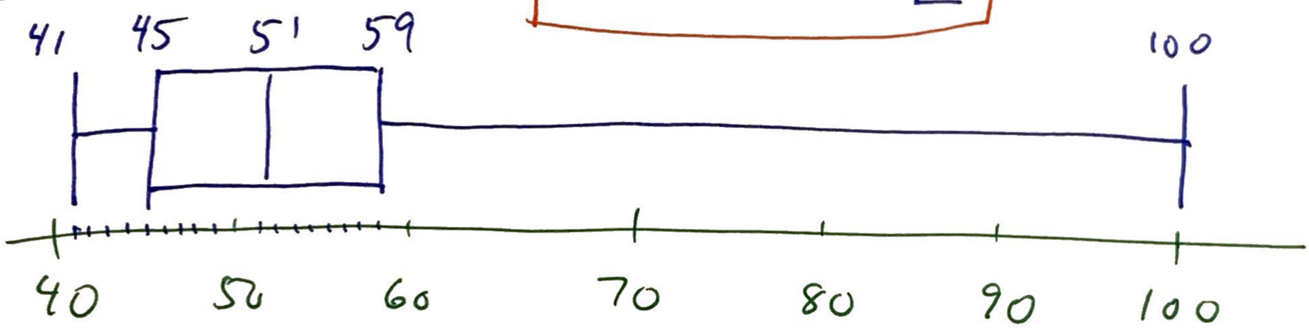
$\frac{51+51}{2} = 51 \leftarrow Q_2$

- **5 number Summary**

min	41
Q_1	45
Q_2	51
Q_3	59
max	100

$IQR = 59 - 45 = \boxed{14}$

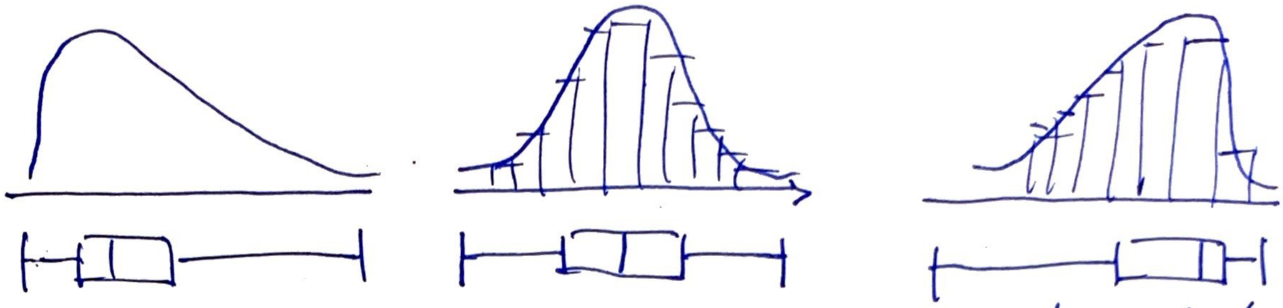
- **Box Plot**



spread of data set is the IQR for non-sym data sets.

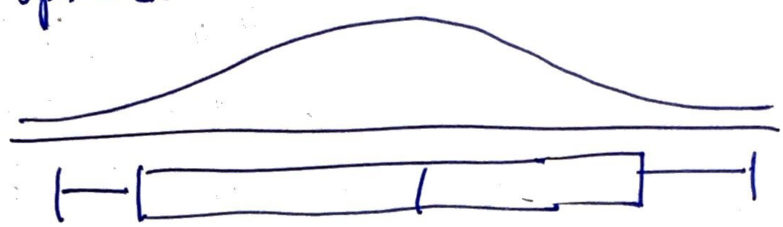
- **Stat disk:** enter data → data → explore data → Box-Plot

Box-Plot vs. Histogram Distribution

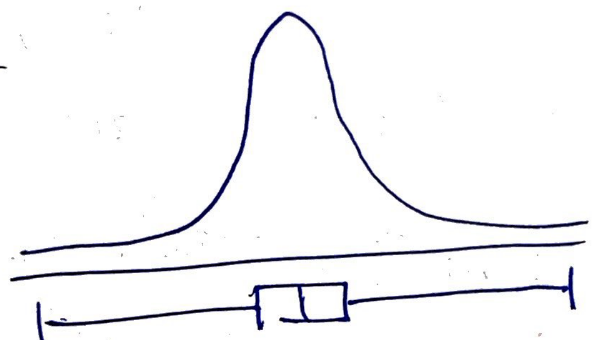


skewed right　　symmetrical　　skewed Left

- wide data spread



- narrow spread



7

Ex Consider the data set (pre-ordered)

2	2	2	2	5	7	8	8	9	9	14	14
14	16	19	20	21	22	22	24	24	27	32	33

$N = 24$

Construct a Dot Plot (see below)

- 5 # summary:

min : 2

Q_1 : $\frac{7+8}{2} = 7.5$

Q_2 : $\frac{14+14}{2} = 14$

Q_3 : $\frac{22+22}{2} = 22$

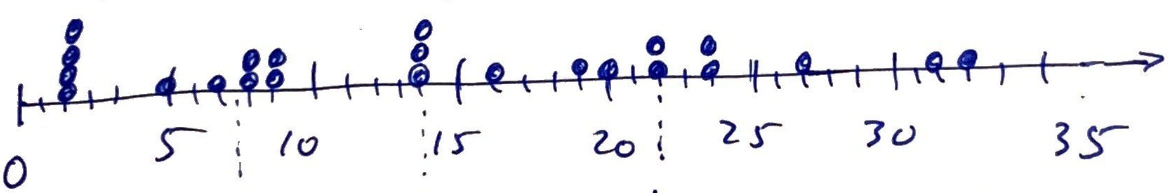
max : 33

- Dot Plot (not necc'y)

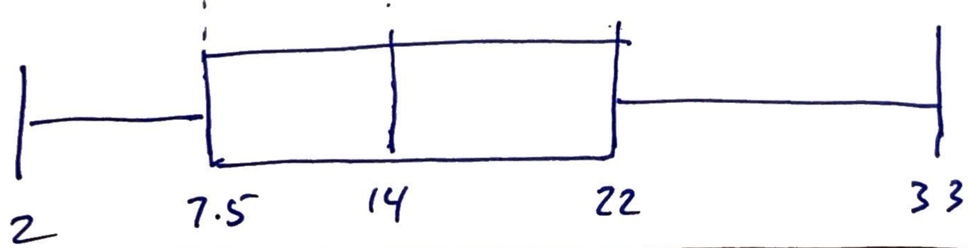
mode = 2

median = 14

- Ruler



- BoxPlot



- statdisk
  - → col in col 1 of data editor
  - → data → box plot → select col 1

→ Box Plot for traditional

→ Modified Box Plot for outlier display

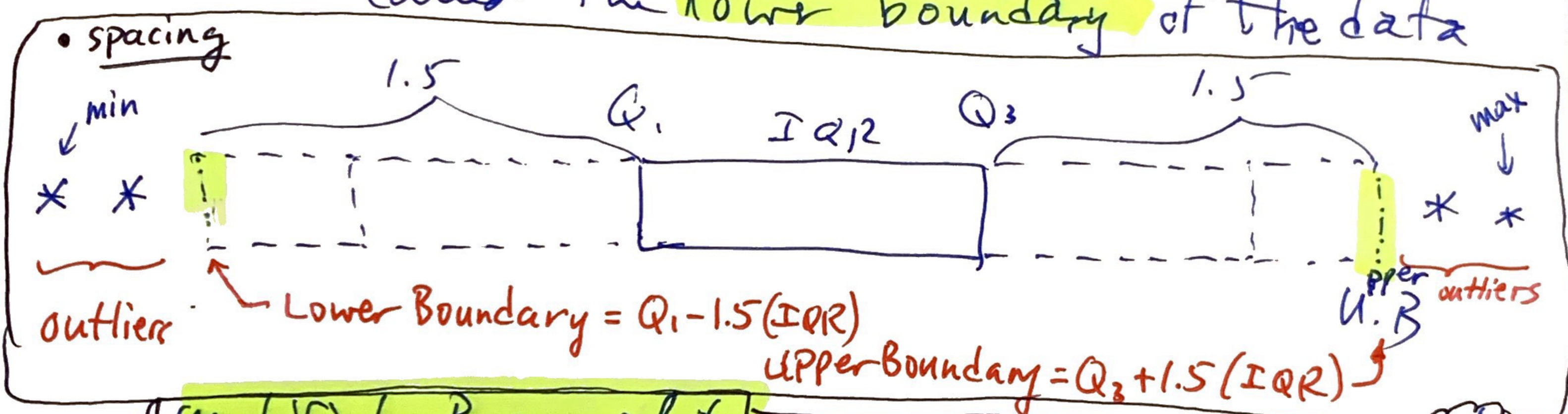
Outliers An outlier is a data point that stands apart from the data set.

→ outliers may be erroneous : Correct or ignore
→ outliers may be a fact : explain or ignore (but tell the reader why).

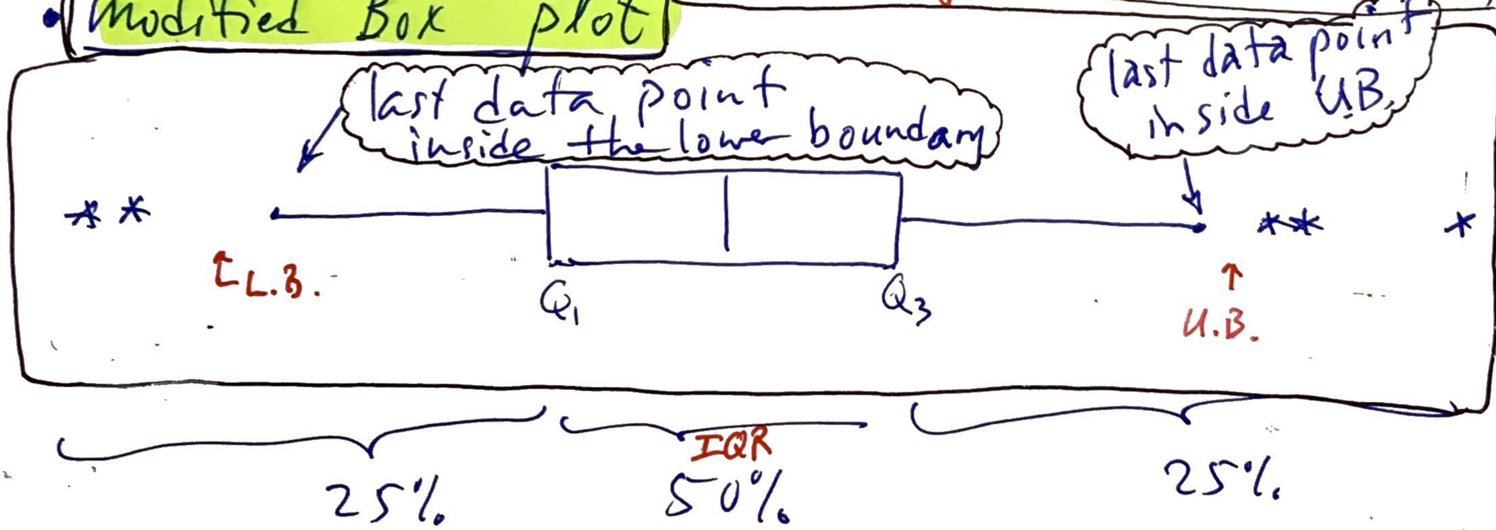
- We identify outliers as being points that are $\boxed{1.5\ IQR\text{'s}}$ above Q_3, this is called the **Upper Boundary**

—OR—

- Points that are $\boxed{1.5\ IQR\text{'s}}$ below Q_1, this is called the **lower boundary** of the data



- **Modified Box plot**



Ex Modified Box Plot

- ordered data

41, 42, 42, 44, 45, 45, 46, 49, 49, 51, 51, 53, 56, 57, 59, 59, 65, 67, 71, 90, 100

Q_1 Q_2

$N = 21$

- 5 # summary

min : 41

Q_1 : $(45+45)/2 = 45$

Q_2 : 51

Q_3 : $(59+65)/2 = 62$

$IQR = 62 - 45 = \boxed{17}$

max : 100

- Lower Boundary: $Q_1 - 1.5\,IQR$

$= 45 - 1.5(17)$

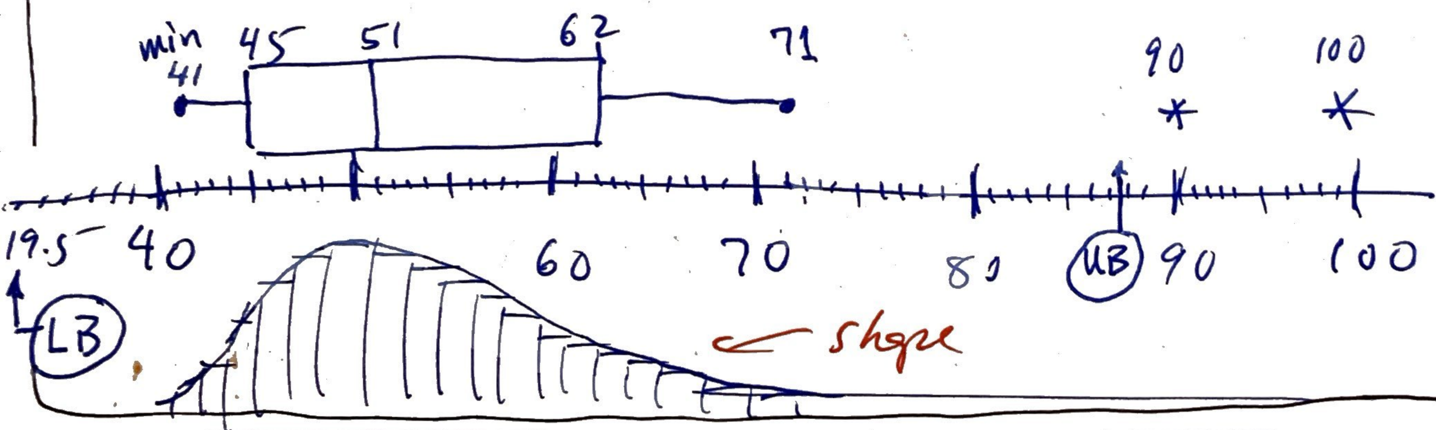
$= \boxed{19.5}$

- Upper Boundary: $Q_3 + 1.5\,IQR$

$= 62 + 1.5(17)$

$= \boxed{87.5}$

- Modified Box Plot w/ ruler



⊛ Percentiles

Quartiles divide the data into quarters.

we may desire more refined details

Ex what data point divides the upper 1/3 from the lower 2/3 ?

- We use percentiles for this task.

Def Given a number p, between 1 and 99, the p^{th} percentile (p^{th} %) separates the lowest p% from the upper (100-p) %

Procedure Finding the p^{th} % (p^{th} percentile)

1. order the data set, count the data
2. Find the Locator L, that point to the data value in the data set.

$$\boxed{L = \left(\frac{p}{100}\right) \cdot n}$$

Locator or, in Comp. Sci, pointer.

3. - If L is a whole # then the p^{th} % is the average of the L and L+1 location
   - If L is not a whole number then round-up to get the Location.

EX Find the 58^{th} percentile from the ordered data set below (11)

① ② ③ ④ ⑤ ⑥ ⑦ ⑧ . . . . ← Locator #'s

2, 2, 2, 2, 5, 7, 8, 8, 9, 9, 14, 14 ← data (12)

N=48

14, 16, 19, 20, 21, 22, 22, 24, 24, 27, 32, 33, (13) (24)

33, 33, 34, $\boxed{35}$, 35, 35, 37, 38, 38, 38, 40
25 26 27 28

40, 40, 41, 42, 46, 47, 48, 48, 48, 48, 48, 49

1. Order data (done)
2. Locator of 58^{th} percentile: $\boxed{L = \left(\frac{58}{100}\right)48 = \underline{27.8}}$
3. Round up to 28
4. Count over from the lowest to the 28^{th} data point: Here the 28^{th} data point is $\boxed{35}$
5. State Results:

"The data value 35 separates the lower 58^{th} % of data from the upper 42%"

((Statdisk has no percentile support))

• graphically (Dot Plot with 58^{th} %)