Example:

Group 1					Group 2			
74	79	77	81		65	77	78	74
68	79	81	76		76	73	71	71
81	80	80	78		86	81	76	89
88	83	79	91		79	78	77	76
79	75	74	73		72	76	75	79

Stem and Leaf Diagrams

stem: the first digit of each observation

leaf: the second digit of each obs.

Stem and Leaf

Stem	Leaf	Count
9	1	1
8	8	1
8		
8		
8	3	1
8	00111	5
7	89999	5
7	67	2
7	445	3
7	3	1
7		
6	8	1

6|8 represents 68

Stem and Leaf

Stem	Leaf	Count
8	9	1
8	6	1
8		
8		
8	1	1
7	8899	4
7	666677	6
7	45	2
7	23	2
7	11	2
6		
6		
6	5	1

6|5 represents 65

Different softwares use different deviding method

Devide leaf for values between 0 - 4 & 5 - 9 e.g: {5, 4, 9, 10, 11, 37, 35, 20, 21, 11, 4, 7, 3, 9, 60, 74}:

05 04 09

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

Frequency tables and histograms

Dot diagrams and stem-and-leaf plots are useful for getting to know a data set, but they are not commonly used in papers and presentations.

1

> A **frequency table** is made by first breaking an interval containing all the data into an appropriate number of smaller intervals of equal length. Then tally marks can be recorded to indicate the number of data points falling into each interval. Finally, frequencies, relative frequencies, and cumulative relative frequencies can be added.

2 3 4

Frequency tables and histograms

A frequency table is made by

- First breaking an interval containing all the data into an apropriate number of smaller intervals of **equal length.**
- Then tally marks can be recorded to indicate the number of data points falling into each interval.
- Finally, add frequency, relative frequency and cumlative relative frequency can be added.

Relative Frequency :

$$\frac{\text{each frequency}}{\text{totall \# of obs}}$$

The choice of interval is arbitrary :

① They're equal in size

② rounded numbers are preferred

Table 3.2
Frequency Table for Laid Gear Thrust Face Runouts

Runout (.0001 in.)	Tally	Frequency	Relative Frequency	Cumulative Relative Frequency
5–8	|||	3	.079 = 3/38	.079
9–12	𝍸 𝍸 𝍸 |||	18	18/38 = .474	.553 = 0.079 + 0.474
13–16	𝍸 𝍸 ||	12	.316 = 12/38	.868 = 0.553 + 0.316
17–20	||||	4	4/38 = .105	.974
21–24		0	0 = 0/38	.974
25–28	|	1	1/38 = .026	1.000
		38	1.000	

always = 1.

always one.

Histogram

After making a frequency table, it is common to use the organization provided by the table to create a histogram.

A **(frequency or relative frequency) histogram** is a kind of bar chart used to portray the shape of a distribution of data points.

Guidelines for making histograms:

- Use intervals of equal length
- Show the entire vertical axis starting at *zero*
- Avoid breaking either axes
- keep a uniform scale for axes (tick marks)
- Center bars of appropriate heights at midpoint of the intervals

Histogram

Example:[Bullet penetration depth, pg. 67]

Sale and Thom compared penetration depths for several types of .45 caliber bullets fired into oak wood from a distance of 15 feet. They recorded the penetration depths (in mm from the target surface to the back of the bullets) for two bullet types.

200 grain jacketed bullets	230 grain jacketed bullets
63.8, 64.65, 59.5, 60.7, 61.3, 61.5, 59.8, 59.1, 62.95, 63.55, 58.65, 71.7, 63.3, 62.65, 67.75, 62.3, 70.4, 64.05, 65, 58	40.5, 38.35, 56, 42.55, 38.35, 27.75, 49.85, 43.6, 38.75, 51.25, 47.9, 48.15, 42.9, 43.85, 37.35, 47.3, 41.15, 51.6, 39.75, 41

20 data points.

Frequency table for 200 grain jacketed bulletss

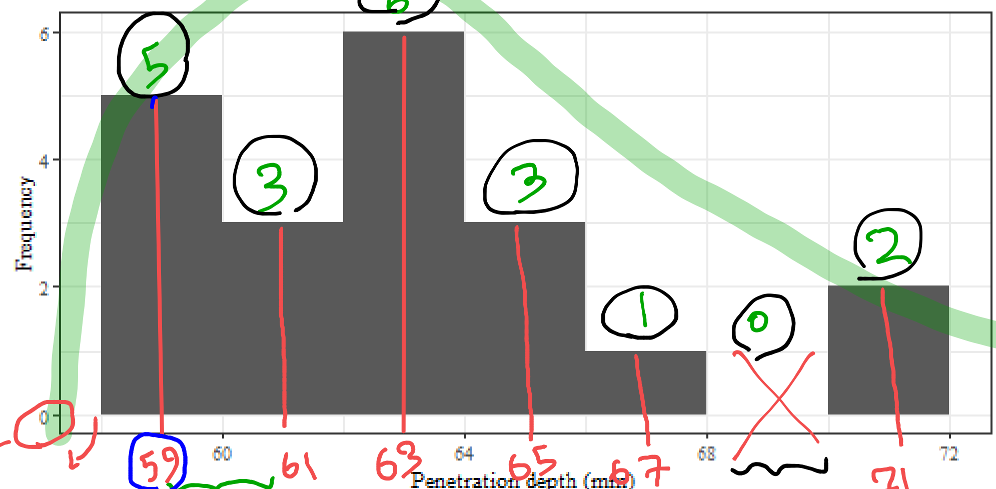
Roundout (0.0001 in)	Center	tally	Frequency	Relative Freq.	Comulative Relative Freq.
58 - 59.99 [58, 60)	59	~~||||~~	5	5/20 = 0.25	.25
60 - 61.99 [60, 62)	61	|||	3	3/20 = .15	.4
62 - 63.99	63	~~||||~~ |	6	6/20 = .3	.7
64 - 65.99	65	|||	3	3/20 = .15	.85
66 - 67.99	67	|	1	1/20 = .05	.9
68 - 69.99	69		0	0/20 = 0	.
70 - 71.99	71	||	2	2/20 = .1	1
			20		

# of data points

in Comulative relative Frequency should always get 1 in the last row.

Histogram

Example:[Bullet penetration depth, pg. 67]



start at zero

Example:[Histogram]

Suppose you have the following data:

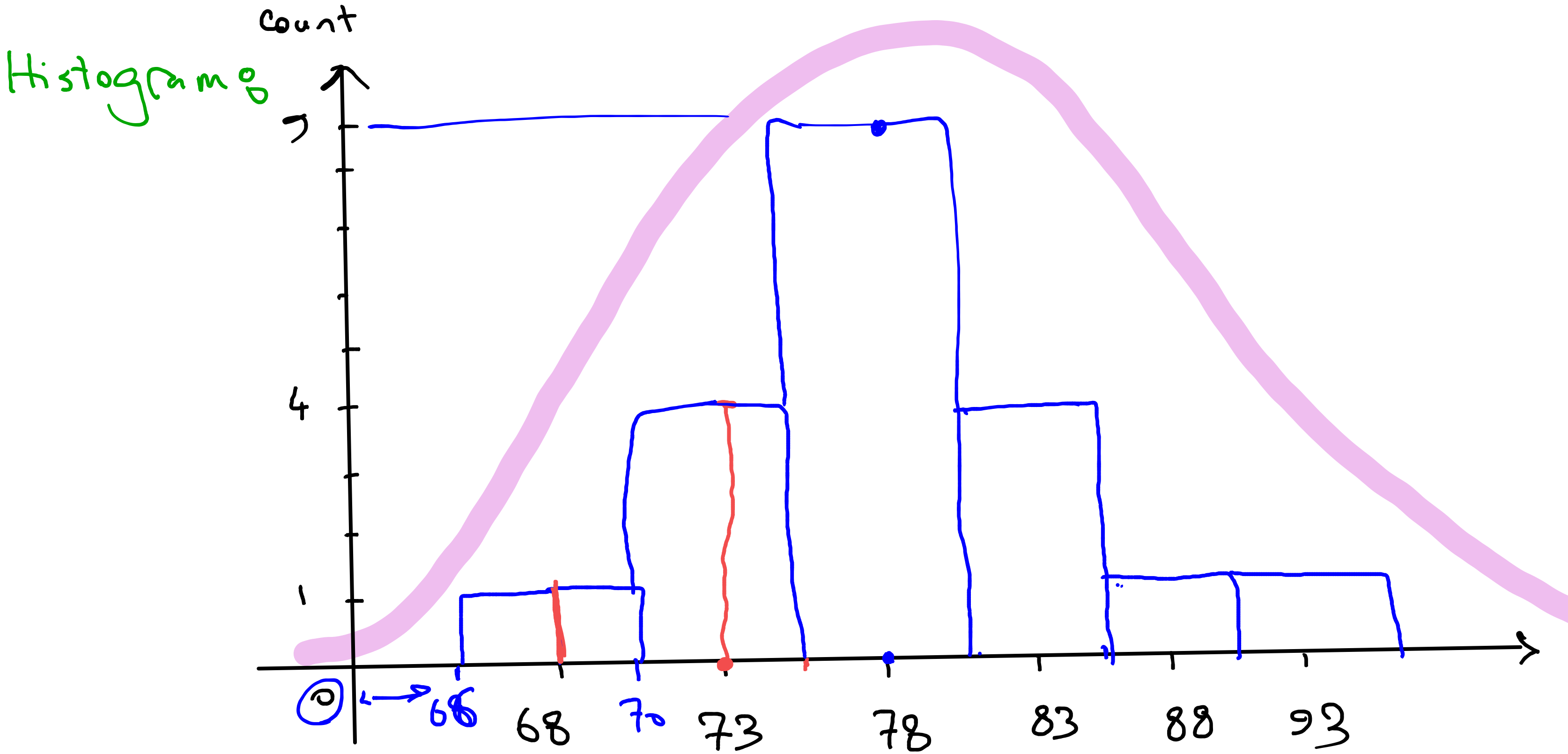
$$74, 79, 77, 81, 68, 79, 81, 76, 81, 80, 80$$
$$, 78, 88, 83, 79, 91, 79, 75, 74, 73$$

Create the corresponding *frequency table* and *frequency histogram*}

use intervals of 4 :

Frequency table :

class	Center Point	tally	Freq.	R.F	C.R.F
66-70	68	|	1	1/20 = .05	.05
71-75	73	////	4	0.2	.25
76-80	78	~~////~~ ////	9	0.45	.7
81-85	83	////	4	0.05	.9
86-90	88	/	1	0.05	1
90-95	93	/	1		
			1/20	1	


Histograms
count
7
4
1
0
66
68
70
73
78
83
88
93

Why plotting data?

Why plotting data?

Why do we plot data?

Information on **location**, **spread**, and **shape** is portrayed clearly in a histogram and can give hints as to the functioning of the physical process that is generating the data.

one mode

• one pick in the center

• unimodal (one pick)

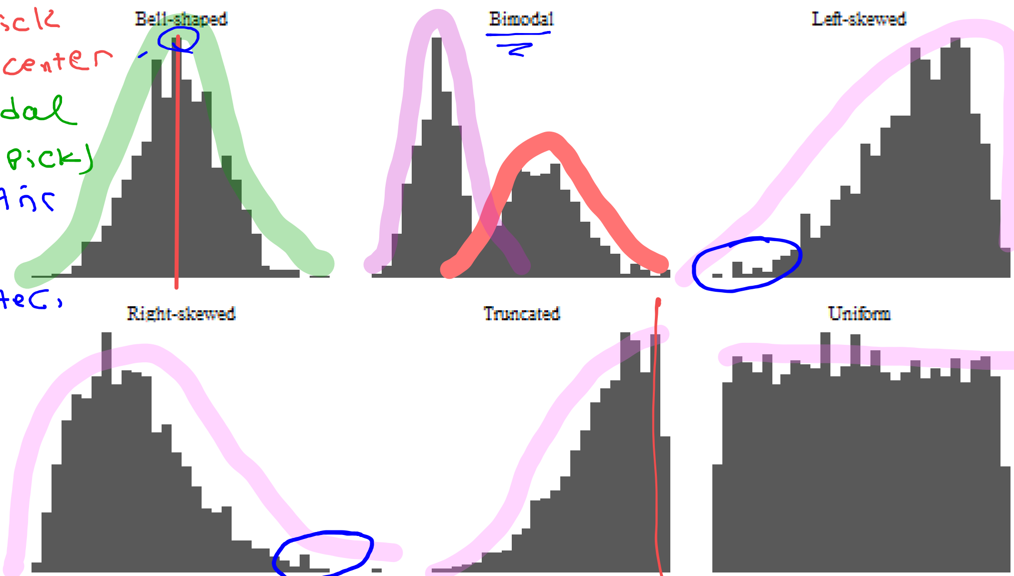
• symmetric around the center

("normal") =

• 2 peaks = 2 modes

• bimodal

long left tail compared to bell-shaped



flat

long right tail compared to bell-shaped

cut off

Why plotting data?

If data on the diameters of machined metal cylinders purchased from a vendor produce a histogram that is decidedly **bimodal**, this suggests

> the machining was done on 2 machines or by two operators or at two different times, etc. ...

If the histogram is **truncated**, this might suggest

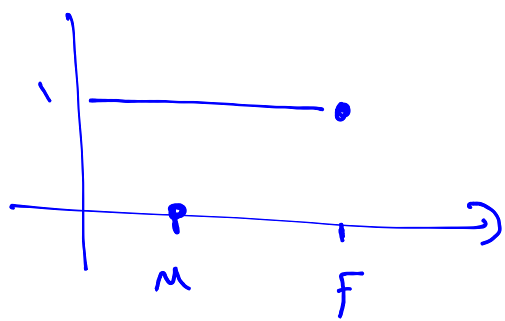
> the cylinders have been 100% inspected

→ The remaining cylinders with excess diameters.

Scatter plots

Dot-diagrams, stem-and-leaf plots, frequency tables, and histograms are univariate tools. But engineering questions often concern multivariate data and *relationships between the quantitative variables*.

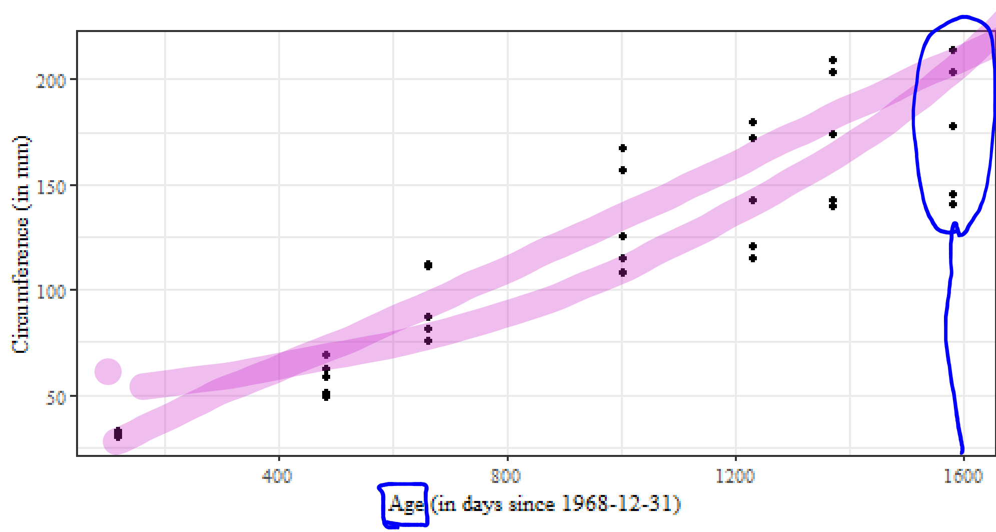
> A **scatterplot** is a simple and effective way of displaying potential relationships between two quantitative variable by assigning each variable to either the x or y axis and plotting the resulting coordinate points.



Scatter plots

Example:[Orange trees]

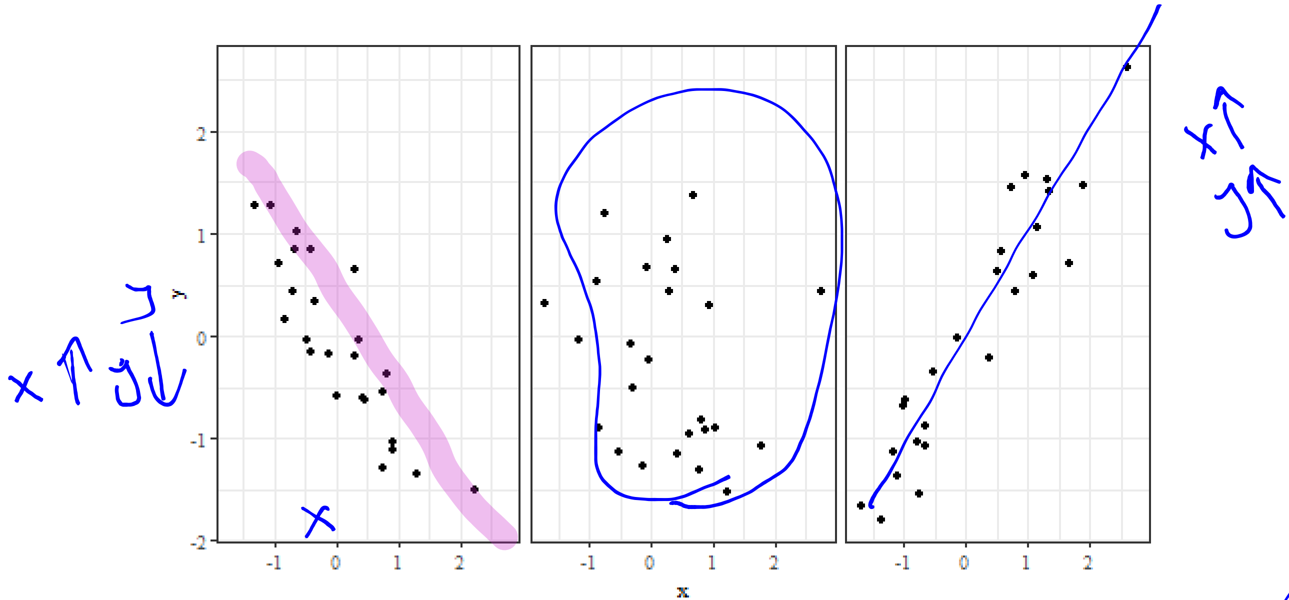
Jim and Jane want to know the relationship between an orange tree's age (in days since 1968-12-31) and its circumference (in mm). They recorded the data for 35 orange trees.



- older trees associated with larger circumferences (positive association/relation)
- older trees have more variability in circumference

Scatter plots

There are three typical association/relationship between two variables:



negatively associated

Non-associated

positively associated

Summaries of Location and Central Tendency

Summaries of Location and Central Tendency

Motivated by asking what is *normal/common/expected* for this data. There are three main types used:

Mean: A "fair" center value. The symbol used differs depending on whether we are dealing with a sample or population:

		Mean
Population		$\mu = \frac{1}{N}\sum_1^N x_i$
Sample		$\bar{x} = \frac{1}{n}\sum_1^n x_i$

N is the population size and **n** is the sample size.

Mode: The most commonly occurring data value in set.

Summaries of Location and Central Tendency

Quantiles: The number that divides our data values so that the proportion, p, of the data values are below the number and the proportion $1 - p$ are above the number.

Median: The value dividing the data values in half (the middle of the values). The median is just the 50th quantile.

Range: The difference between the highest and lowest values (Range = max - min)

IQR: The Interquartile Range, how spread out is the middle 50% (IQR = Q3 - Q1)

Summaries of Location and Central Tendency

y_2

Group 1	Group 2
74 79 77 81	65 77 78 74
68 79 81 76	76 73 71 71
81 80 80 78	86 81 76 89
88 83 79 91	79 78 77 76
79 75 74 73	72 76 75 79

$x_1 \leftarrow$ 74 $\quad$ 73 $\rightarrow x_{20}$

Calculating Mean Think of it as an equal division of the total

- each value in the data is an "x_i" (i is a **subscript**)
- Group 1: $x_1 = 74, x_2 = 79, \ldots, x_{20} = 73$
- The sum: $x_1 + x_2 + x_3 + \ldots + x_{20}$
- divides : $(x_1 + x_2 + x_3 + \ldots + x_{20})/20$
- Or using summation notation: $\frac{1}{20}\sum_{i=1}^{20} x_i$

Quantiles

The Quantile Function

Two useful pieces of notation:

floor: $\lfloor x \rfloor$ is the largest integer smaller than or equal to x

ceiling: $\lceil x \rceil$ is the smallest integer larger than or equal to x

Examples

- $\lfloor 55.2 \rfloor = 55$
- $\lceil 55.2 \rceil = 56$
- $\lfloor 19 \rfloor = 19$
- $\lceil 19 \rceil = 19$
- $\lceil -3.2 \rceil = -3$
- $\lfloor -3.2 \rfloor = -4$

Summaries of Location and Central Tendency

Quantiles

Already familiar with the concept of "percentile".

The p^{th} percentile of a data set is a number greater than p % of the data and less than the rest.

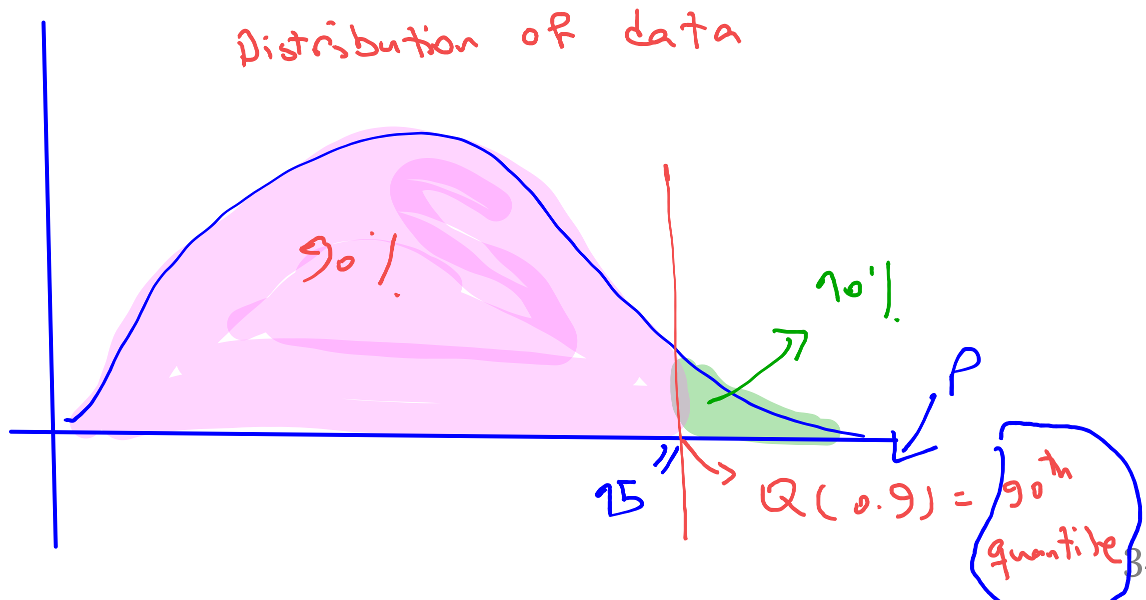
> "You scored at the 90^{th} percentile on the SAT" means that your score was higher than 90% of the students who took the test and lower than the other 10%
>
> "Zorbit was positioned at the 80^{th} percentile of the list of fastest growing companies compiled by INC magazine.” means Zorbit was growing faster than 80% of the companies in the list and slower than the other 20%.

Summaries of Location and Central Tendency

Quantiles

- It is more convenient to work in terms of fractions between 0 and 1 rather than percentages between 0 and 100. We then use terminology **Quantiles** rather than percentiles.
- For a number **p** between 0 and 1, the **p quantle** of a distribution is a number such that a fraction p of the distribution lies to the left of that value, and a fraction 1-p of the distribution lies to the right.

e.g



Summaries of Location and Central Tendency

Quantiles

For a data set consisting of n values that when ordered are $x_1 \leq x_2 \leq \cdots \leq x_n$,

- if $p = \frac{i-.5}{n}$ for a positive integer $i \leq n$, the p **quantile** of the data set is

$$Q(p) = Q\left(\frac{i-.5}{n}\right) = x_i$$

(the ith smallest data point will be called the $\frac{i-.5}{n}$ quantile)

Summaries of Location and Central Tendency

Quantiles

- for any number p between $\frac{.5}{n}$ and $\frac{n-.5}{n}$ that is not of the form $\frac{i-.5}{n}$ for an integer i, the p **quantile** of the data set will be obtained by linear interpolation between the two values of $Q\left(\frac{i-.5}{n}\right)$ with corresponding $\frac{i-.5}{n}$ that bracket p.

In both cases, the notation $Q(p)$ will denote the p quantile.

Summaries of Location and Central Tendency

The Quantile Function

For a data set consisting of n values that when ordered are $x_1 \leq x_2 \leq \ldots \leq x_n$ and $0 \leq p \leq 1$.

We define the **quantile function** $Q(p)$ as:

$$Q(p) = \begin{cases} x_i & \lfloor n \cdot p + .5 \rfloor = n \cdot p + .5 \\ x_i + (np - i + .5)(x_{i+1} - x_i) & \lfloor n \cdot p + .5 \rfloor \neq n \cdot p + .5 \end{cases}$$

(note: this is the definition used in the book - it's just written using *floor* and *ceiling* instead of in words)

Quantile Function:

i

$$Q(P) = \begin{cases} x_i & \text{if } \lfloor n.P + 0.5 \rfloor = n.P + 0.5 \\ x_i + (n.P - i + 0.5)(x_{i+1} - x_i) & \text{if } \lfloor nP + 0.5 \rfloor \neq nP + 0.5 \end{cases}$$

$x_1 =$ $= x_n$

- $Q\left(\frac{1-.5}{n}\right)$ is called the **minimum** and $Q\left(\frac{n-.5}{n}\right)$ is called the **maximum** of a distribution.
- $Q(.5)$ is called the **median** of a distribution. $Q(.25)$ and $Q(.75)$ are called the **first (or lower) quartile** and **third (or upper) quartile** of a distribution, respectively.
- The **interquartile range (IQR)** is defined as

$$IQR = Q(.75) - Q(.25)$$

.

- An **outlier** is a data point that is larger than $Q(.75) + 1.5 * IQR$ or smaller than $Q(.25) - 1.5 * IQR$.

Example: Find the median, first quartile, 17th quantile and 65th quantile for the following set of data values:

$$58, 76, 66, 61, 50, 77, 67, 64, 41, 61$$

First notice that $n = 10$. It is possible helpful to set up the following table:

- **Step 1: sort the data**

data	41	50	58	61	61	64	66	67	76	77
i	1	2	3	4	5	6	7	8	9	10

Example: Find the median, first quartile, 17th quantile and 65th quantile for the following set of data values:

58, 76, 66, 61, 50, 77, 67, 64, 41, 61

- **Step 2: find** $\frac{i-.5}{n}$

$n = 10$

$Q(p)$

data	41	50	58	61	61	64	66	67	76	77
i	1	2	3	4	5	6	7	8	9	10
$p = \frac{i-.5}{n}$	0.05	0.15	0.25	0.35	0.45	0.55	0.65	0.75	0.85	0.95

- **Step 3: find $Q(p)$**

data	41	50	58	61	61	64	66	67	76	77
i	1	2	3	4	5	6	7	8	9	10
$\frac{i-.5}{n}$	0.05	0.15	0.25	0.35	0.45	0.55	0.65	0.75	0.85	0.95

$$Q(p) = \begin{cases} x_i & \lfloor n \cdot p + .5 \rfloor = n \cdot p + .5 \\ x_i + (np - i + .5)\,(x_{i+1} - x_i) & \lfloor n \cdot p + .5 \rfloor \neq n \cdot p + .5 \end{cases}$$

Finding the first **quartile** ($Q(.25)$):

- $np + .5 = 10 \cdot .25 + .5 = 3. = i$
- since $\lfloor 3 \rfloor = 3$

  then $\boxed{i = 3}$ and

  $Q(.25) = x_3 = 58$

Your turn

Find the median

data	41	50	58	61	61	64	66	67	76	77
i	1	2	3	4	5	6	7	8	9	10
$\frac{i-.5}{n}$	0.05	0.15	0.25	0.35	0.45	0.55	0.65	0.75	0.85	0.95

$$Q(p) = \begin{cases} x_i & \lfloor n \cdot p + .5 \rfloor = n \cdot p + .5 \\ x_i + (np - i + .5)(x_{i+1} - x_i) & \boxed{\lfloor n \cdot p + .5 \rfloor} \neq n \cdot p + .5 \end{cases}$$

$p = 0.5$

- $np + .5 = 10 \cdot 0.5 + \boxed{0.5} = \boxed{5.5}$
- since $\lfloor 5.5 \rfloor = 5$ then $i = 5$ and

$$Q(.5) = x_i + (n \cdot p - i + .5) \cdot (x_{i+1} - x_i)$$

$$= x_5 + (10 \cdot 0.5 - 5 + .5) \cdot (x_{5+1} - x_5)$$

$$= x_5 + (.5) \cdot (x_6 - x_5)$$

$$= 61 + (.5) \cdot (64 - 61)$$

$$\boxed{= 62.5}$$

Finding $Q(.17)$

- $np + .5 = 10 \cdot 0.17 + 0.5 = 2.2.$
- since $\lfloor 2.2 \rfloor = 2$ then $\boxed{i = 2}$ and

$$Q(.17) = x_i + (n \cdot p - i + .5) \cdot (x_{i+1} - x_i)$$

$$= x_2 + (10 \cdot 0.17 - 2 + .5) \cdot (x_{2+1} - x_2)$$

$$= x_9 + (.2) \cdot (x_3 - x_2)$$

$$= 50 + (.2) \cdot (58 - 50)$$

$$= 51.6$$

$= p$

Finding $Q(.65)$

- $np + .5 = 10 \cdot 0.65 + 0.5 = 7.$
- since $\lfloor 7 \rfloor = 7$ then $i = 7$ and

$$
\begin{aligned}
Q(.65) &= x_i + (n \cdot p - i + .5) \cdot (x_{i+1} - x_i) \\
&= x_7 + (10 \cdot 0.65 - 7 + .5) \cdot (x_{7+1} - x_7) \\
&= x_7 + (0) \cdot (x_8 - x_7) \\
&= x_7 + 0 \\
&= 66
\end{aligned}
$$

Section 3.2: Plots

Boxplots

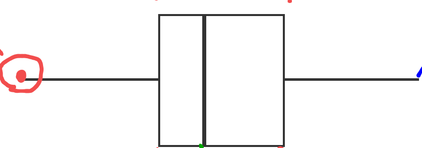
Boxplots

Quantiles are useful in making *boxplots*, an alternative to dot diagrams or histograms.

The boxplot shows less information, but many can be placed side by side on a single page for comparisons.

Smallest data point bigger than or equal to $Q(0.25) - 1.5 * IQR$

$1.5 * IQR$ IQR $1.5 * IQR$



largest data point less than or equal to $Q(0.75) + 1.5 * IQR$

$Q(0.25)$ $Q(0.5)$ $Q(0.75)$

any points NOT in the interval

$[Q(.25) - 1.5 * IQR,\ Q(.75) + 1.5 * IQR]$

are plotted separately.

Boxplots

A simple plot making use of the first, second and third quartiles (i.e., $Q(.25)$, $Q(.5)$ and $Q(.75)$.

1. A box is drawn so that it covers the range from $Q(.25)$ up to $Q(.75)$ with a vertical line at the median.
2. Whiskers extend from the sides of the box to the furthest points within 1.5 IQR of the box edges
3. Any points beyond the whiskers are plotted on their own.

Example: Draw boxplots for the groups using quantile function

```
   Group 1          Group 2
74 79 77 81      65 77 78 74
68 79 81 76      76 73 71 71
81 80 80 78      86 81 76 89
88 83 79 91      79 78 77 76
79 75 74 73      72 76 75 79
```

solution: First we need the quartile values:

	$Q(.25)$	$Q(.5)$	$Q(.75)$
Group 1	75.5	79	81
Group 2	73.5	76	78.5

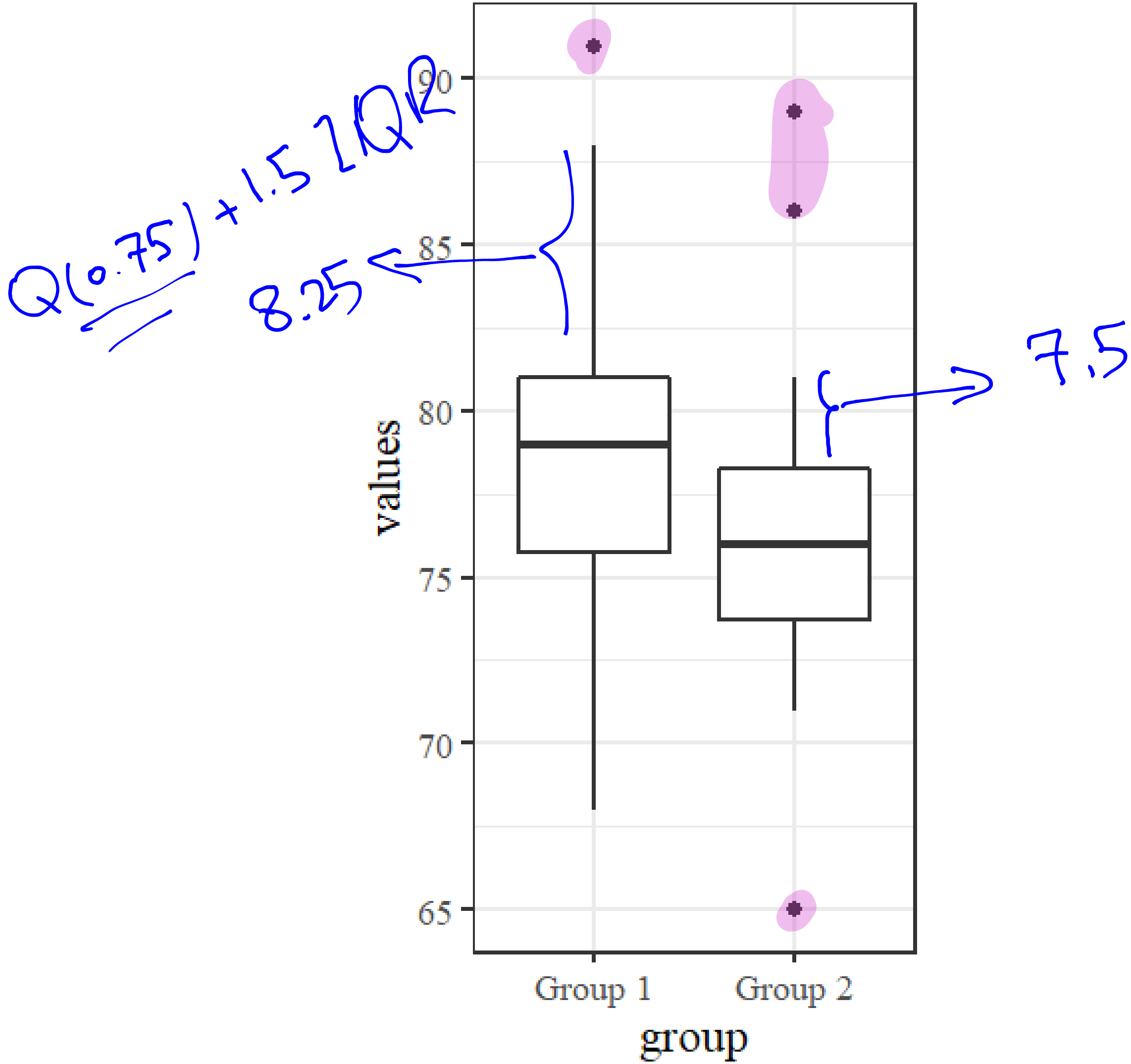
This means that Group 1 has IQR = 5.5 and

- 1.5*IQR = 8.25

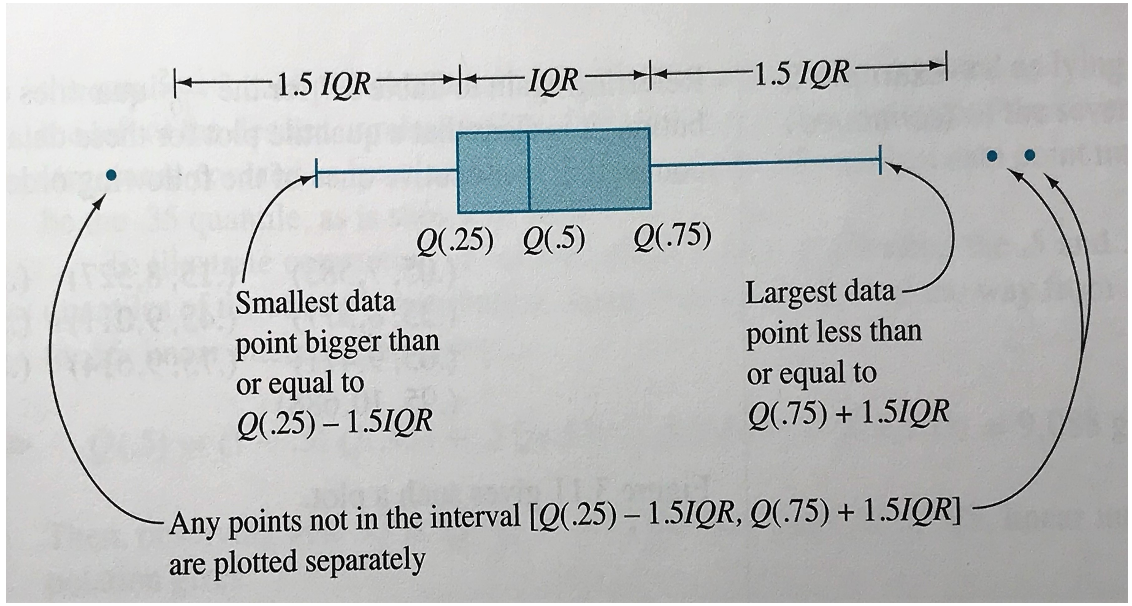
while Group 2 has IQR = 5 and

- 1.5*IQR = 7.5

Example:



Anatomy of a Boxplot



Example:[Bullet penetration depths, cont'd]

i	` $\frac{i-.5}{20}$ `	200 grain bullets	230 grain bullets
1	0.025	58.00	27.75
2	0.075	58.65	37.35
3	0.125	59.10	38.35
4	0.175	59.50	38.35
5	0.225	59.80	38.75
6	0.275	60.70	39.75
7	0.325	61.30	40.50
8	0.375	61.50	41.00
9	0.425	62.30	41.15
10	0.475	62.65	42.55
11	0.525	62.95	42.90
12	0.575	63.30	43.60
13	0.625	63.55	43.85
14	0.675	63.80	47.30
15	0.725	64.05	47.90
16	0.775	64.65	48.15
17	0.825	65.00	49.85
18	0.875	67.75	51.25
19	0.925	70.40	51.60
20	0.975	71.70	56.00

For 230:

$$Q(.25) = X_5 + (np - 5 + 0.5)(X_6 - X_5) = 39.25 \text{ mm}$$

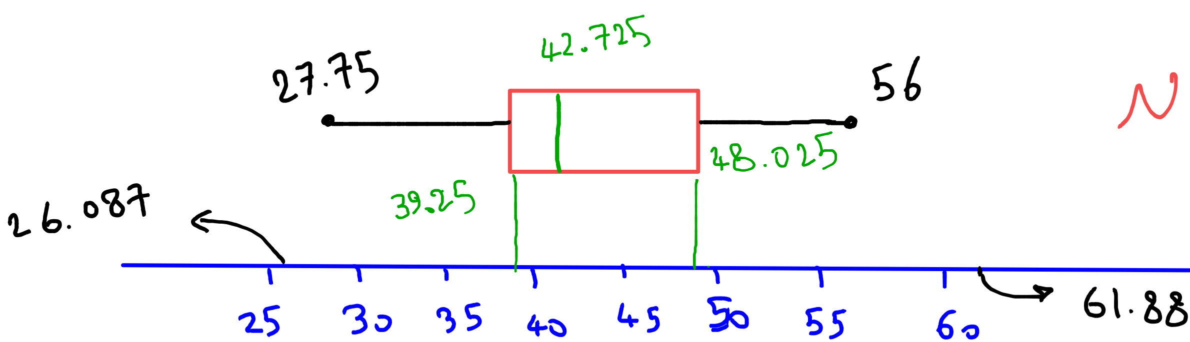
$$Q(.5) = X_{10} + (np - 10 + 0.5)(X_{11} - X_{10}) = 42.725 \text{ mm}$$

$$Q(.75) = X_{15} + (np - 15 + 0.5)(X_{16} - X_{15}) = 48.025 \text{ mm}$$

$$IQR = Q(.75) - Q(0.25) = 48.025 - 39.25 = 8.775$$

$$1.5 * IQR = 13.163$$

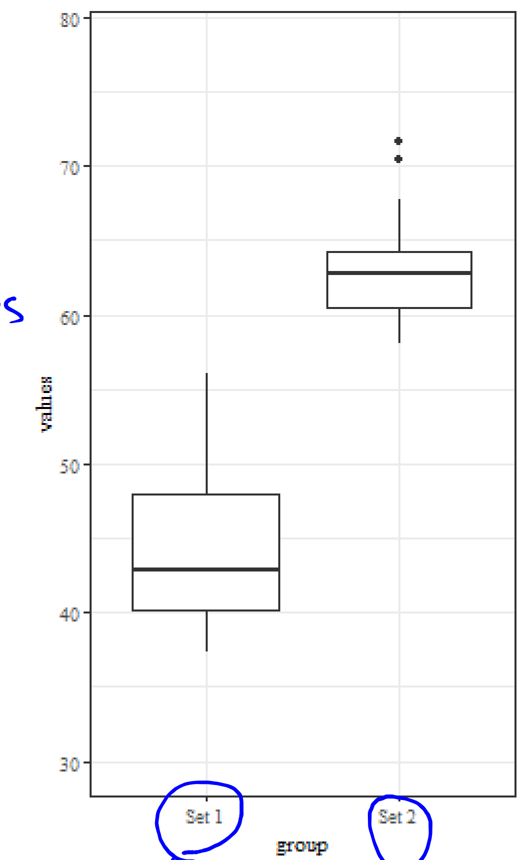
$$\begin{cases} Q(.75) + 1.5 * IQR = 61.88 \text{ mm} \\ Q(.25) - 1.5 * IQR = 26.087 \text{ mm} \end{cases}$$



No outliers!

Example:[Bullet penetration depths, cont'd]

side-by-side Boxplots for the two types of bullets.



230 grain

200 grain