Week 3

Monday, 6 November 2023 9:19 AM

Frequency tables → cells representing counts

For example, the counts of passengers on the *Titanic* are shown for each passenger class.

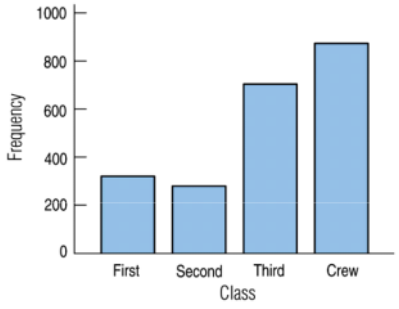
Class	Count
First	325
Second	285
Third	706
Crew	885
Total	2201

Relative frequency tables → gives percentages

Class	%
First	14.77
Second	12.95
Third	32.08
Crew	40.21

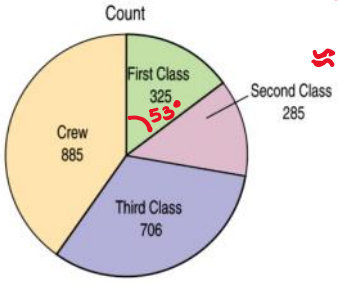
$\frac{325}{2201} \times 100 = 14.766$

BAR CHARTS



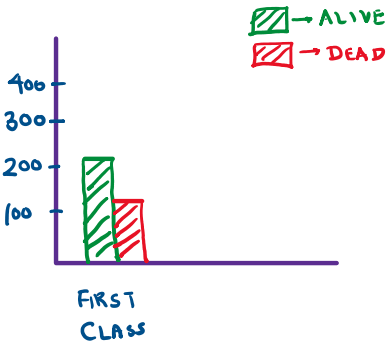
PIE CHARTS

$\frac{14.77}{100} \times 360^\circ \approx 53^\circ$



Contingency table → Two categorical variables

		Class				
		First	Second	Third	Crew	Total
Survival	Alive	203	118	178	212	711
	Dead	122	167	528	673	1490
	Total	325	285	706	885	2201



Question 1 : Titanic Contingency table calculations activity

Survival	Class: First	Second	Third	Crew	Total
Alive	203	118	178	212	711
Dead	122	167	528	673	1490
Total	325	285	706	885	2201

(a) $\frac{711}{2201} \times 100 = 32.303\%$

(b) $\frac{325}{2201} \times 100 = 14.766\%$

(c) $\frac{203}{711} \times 100 = 28.551\%$

(d) $\frac{203}{325} \times 100 = 62.461\%$

a) What proportion of the passengers survived?
b) What proportion of the passengers were in first class?
c) What proportion of the surviving passengers were in first class?
d) What proportion of the first class passengers survived?
e) What proportion of the passengers were in first class and survived?

$\frac{203}{2201} \times 100 = 9.22\%$ → 9.22

NUMERICAL SUMMARIES → Quantitative / Numerical data

- MEASURE OF CENTRE
  - MEAN
  - MEDIAN
  - MODE
- MEASURE OF SPREAD
  - RANGE
  - IQR → Interquartile range
  - SD → Standard deviation

MEAN OR AVERAGE = $\frac{\text{Add all the observations}}{\text{Number of observations}}$

For example, for the children's ages 10, 7, 8, 7 and 4, the mean is

$$\bar{y} = \frac{10+7+8+7+4}{5} = \frac{36}{5} = 7.2$$

MEDIAN → Middle value when data are ordered

4, 7, (7), 8, 10

Median is 7

MODE → most common value

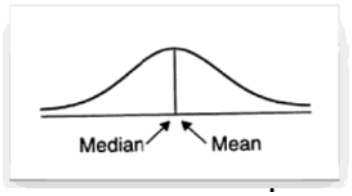
Mode is 7

Median example:

Median example:

E.g. for 4 4 **8** 9 11, Median is 8

E.g. for 4 4 6 | 8 9 11 Median is 7



SYMMETRIC
MEAN

POSITIVELY SKEWED
MEDIAN

NEGATIVELY SKEWED
MEDIAN

MEASURE OF SPREAD

range = max – min

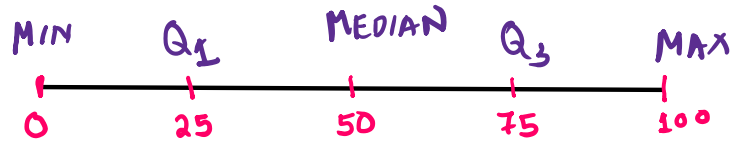
= 10 – 4 = 6

MIN 4, 7, 7, 8, 10 MAX

IQR → Inter quartile Range

$IQR = Q_3 - Q_1$

Upper quartile – Lower Quartile



4, 7, 7, 8, 10

MIN MEDIAN MAX

$IQR = Q_3 - Q_1$

$= 8 - 7$

$= 1$

Standard deviation →

$$s = \sqrt{\frac{\sum (y - \bar{y})^2}{n-1}}$$

$\bar{y} = 7.2$

y	$y - \bar{y}$	$(y - \bar{y})^2$
4	-3.2	10.24
7	-0.2	0.04
7	-0.2	0.04
8	0.8	0.64
10	2.8	7.84

$\bar{y} = 7.2$

$n = 5$

8	0.8	
10	2.8	7.84
		18.80

$$s = \sqrt{\frac{18.80}{5-1}} = 2.16$$

Question 2 : Summary statistics exercise

Lower Fence — Upper fence

For the values: 6 7 8 8 (9) 10 10 11 18,

Median = 9

Mode = 8, 10

find the mean, median, mode, standard deviation, range and IQR.

Are there any outliers?

Using Descriptive Statistics Calculator:

https://www.calculatorsoup.com/calculators/statistics/descriptivestatistics.php

Descriptive Statistics Calculator

Enter Data Set

6, 7, 8, 8, 9, 10, 10, 11, 18

(•) Sample ( ) Population

Clear | Calculate

Answer:

Descriptive Statistics:

Minimum	min =	6
Maximum	max =	18
Range	R =	12
Size	n =	9
Sum	sum =	87
Mean	$\bar{x}$ =	9.66666667
Median	$\tilde{x}$ =	9
Mode	mode =	8, 10

Standard Deviation	s =	3.5
Variance	s^2 =	12.25
Mid Range	MR =	12
Quartiles		Quartiles: Q_1 --> 7.5 ✓ Q_2 --> 9 Q_3 --> 10.5 ✓
Interquartile Range	IQR =	3 ✓
Outliers		18

Maximum − Minimum

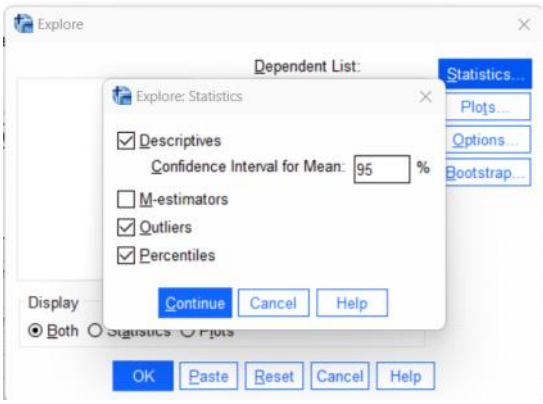
Lower Fence = Q1−1.5×IQR = 7.5−(1.5×3) = 3

Upper Fence = Q3+1.5×IQR = 10.5+(1.5×3) = 15

Using SPSS,

→ Type the values in "Data View"

→ Analyze → Descriptive statistics → Explore



Descriptives

			Statistic	Std. Error
VAR00001	Mean		9.6667	1.16667
	95% Confidence Interval for Mean	Lower Bound	6.9763	
		Upper Bound	12.3570	
	5% Trimmed Mean		9.4074	
	Median		9.0000	
	Variance		12.250	
	Std. Deviation		3.50000	
	Minimum		6.00	
	Maximum		18.00	
	Range		12.00	
	Interquartile Range		3.00	
	Skewness		1.887	.717
	Kurtosis		4.483	1.400

Mean > Median

9.6667 > 9.000

Skewed to the Right

Percentiles

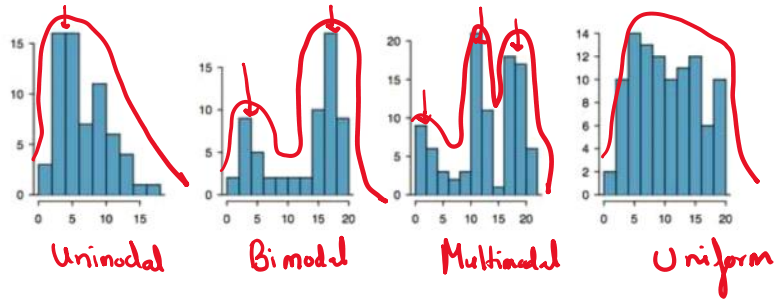
		Percentiles						
		5	10	25	50	75	90	95
Weighted Average (Definition 1)	VAR00001	6.0000	6.0000	7.5000	9.0000	10.5000	.	.
Tukey's Hinges	VAR00001			8.0000	9.0000	10.0000		

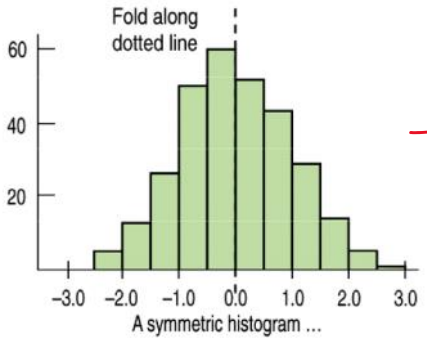
Q_1 Median Q_3

Charts for Quantitative Data: HISTOGRAM & BOX PLOTS

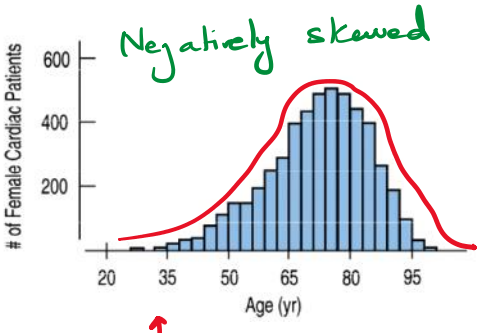
Does the histogram have a single, central hump or several separated peaks?

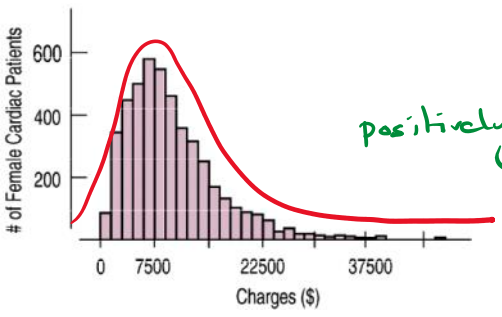
- Humps in a histogram are called modes.
- A histogram with one main peak is called unimodal;
- A histogram with two peaks are bimodal;
- A histogram with three or more peaks are called multimodal.

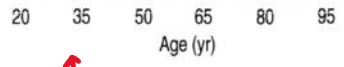




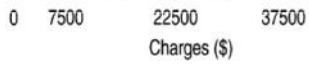
→ Unimodal and symmetric







Unimodal and skewed to the Left



Unimodal and skewed to the Right

Five number summaries and boxplots

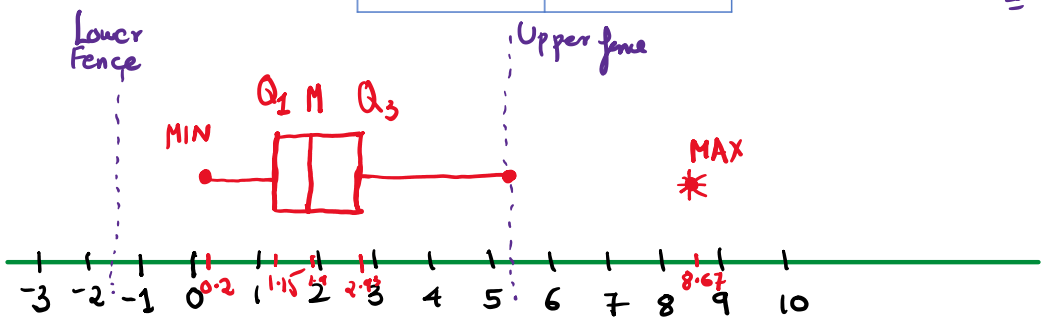
$IQR = Q_3 - Q_1 = 2.93 - 1.15 = 1.78$

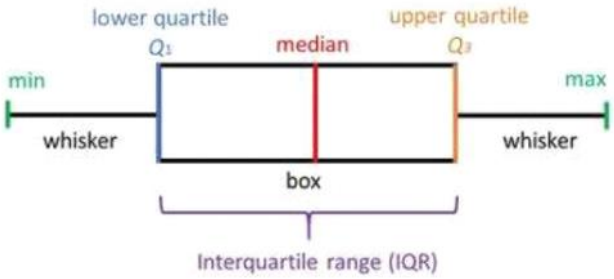
- The five-number summary of a distribution reports its median, quartiles, and extremes (maximum and minimum).
  - Example: The five-number summary for the daily wind speed is:

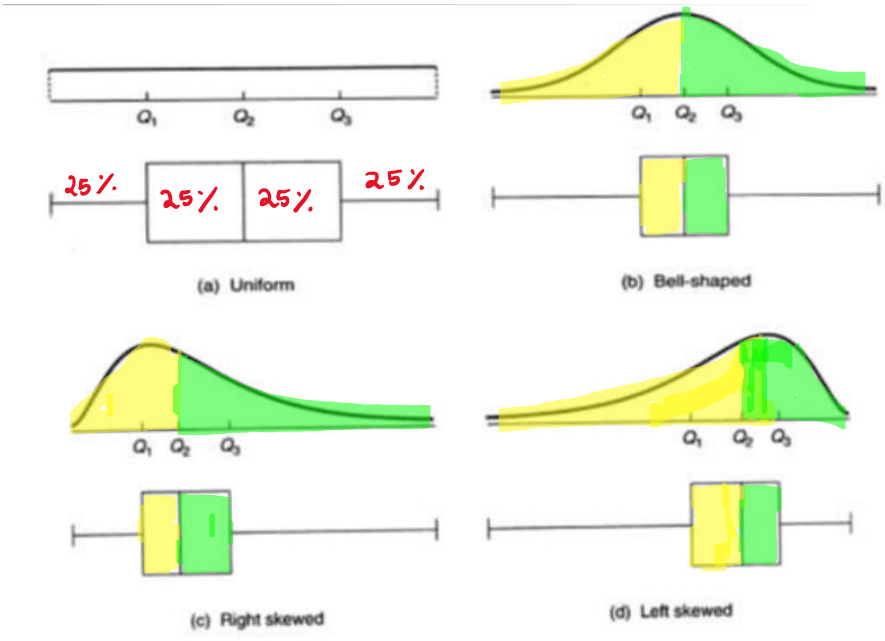
Max	8.67
Q3	2.93
Median	1.9
Q1	1.15
Min	0.2

Lower fence = Q1 − 1.5×IQR
= 1.15 − (1.5 × 1.78)
= −1.52

Upper fence = Q3 + 1.5×IQR
= 2.93 + (1.5 × 1.78)
= 5.6







Unimodal and symmetric