CS1512

CS2013 Mathematics for Computing Science

Kees van Deemter

Probability and statistics

UNIVERSITY OF ABERDEEN

What this is going to be about

1. Suppose the statement p is true
   and the statement q is true.
   What can you say about the statement p and q ?

What this is going to be about

1. Suppose the statement p is true
   and the statement q is true.
   What can you say about the statement p and q ?
   In this case, p and q is also true.
2. Suppose the statement p has a probability of .5
   and the statement q has a probability of .5.
   What can you say about the statement p and q ?

What this is going to be about

1. Suppose the statement p is true
   and the statement q is true.
   What can you say about the statement p and q ?
   In this case, p and q is also true.
2. Suppose the statement p has a probability of .5
   and the statement q has a probability of .5.
   What can you say about the statement p and q ?

It depends! If p and q are independent then p and q has a probability of .25 But suppose

p = It will snow (some time) tomorrow and
q = It will be below zero (some time) tomorrow

Then p and q has a probability >.25

What this is going to be about

1. Suppose the statement p is true
   and the statement q is true.
   What can you say about the statement p and q ?
   In this case, p and q is also true.
2. Suppose the statement p has a probability of .5
   and the statement q has a probability of .5.
   What can you say about the statement p and q ?

It depends! If p and q are independent then p and q has a probability of .25 But suppose

p = It will snow (sometime) tomorrow and
q = It will **not** snow (any time) tomorrow

Then p and q has a probability 0

Before we get there ...

Some basic concepts in statistics

- different kinds of data
- ways of representing data
- ways of summarising data

Useful in CS. For example to

- assess whether a computer simulation is accurate
- assess whether one user interface is more user friendly than another
- estimate the expected run time of a program (on typical data)

Lecture slides on statistics
are based on originals by Jim Hunter.

Sources

Text book (parts of chapters 1-6):

Essential Statistics (Fourth Edition)
D.G.Rees
Chapman and Hall
2001
(Blackwells, ~£28)

Some definitions

Sample space (population)

- Set of entities of interest, also called elements
- this set may be infinite
- entities can be physical objects, events, etc. ...

Sample

- subset of the sample space

More definitions

Variable

- an attribute of an element which has a value (e.g., its height, weight, etc.)

Observation

- the value of a variable as recorded for a particular element
- an element will have variables with values but they are not observations until we record it

Sample data

- set of observations derived from a sample

Descriptive and Inferential Statistics

Descriptive statistics:

- Summarising the sample data (as a number, graphic ...)

Inferential statistics:

- Using data from a sample to infer properties of the sample space
- Chose a ‘representative sample’
  (properties of sample match those of sample space – difficult)
- In practice, use a ‘random sample’
  (each element has the same likelihood of being chosen)

Variable types

Qualitative:

- Nominal/Categorical (no ordering in values)
  - e.g. sex, occupation
- Ordinal (ranked)
  - e.g. class of degree (1, 2.1, 2.2,...)

Quantitative:

- Discrete (countable) – [integer]
  - e.g. number of people in a room
- Continuous – [double]
  - e.g. height

Examples

1. A person’s marital status
2. The length of a CD
3. The size of a litter of piglets
4. The temperature in degrees centigrade

Examples

1. A person’s marital status

   Nominal/categorical

2. The length of a CD

   Quantitative; continuous or discrete?

   This depends on how you model length (minutes or bits)

3. The size of a litter of piglets

   Quantitative, discrete (if we mean the number of pigs)

4. The temperature in degrees centigrade

   Quantitative, continuous

   (Even though it does not make sense to say that 20^0 is twice as warm as 10^0)

Footnote: We us the term `Continuous` loosely: For us a variable is continuous/dense (as opposed to discrete) if between any values x and y, there lies a third value z.

Summarising data

Categorical (one variable):

- X is a categorical variable with values: $a_1, \; a_2, \; a_3, \; \ldots \; a_k, \; \ldots \; a_K$

  $(k = 1, 2, 3, \ldots K)$

- f_k = number of times that a_k appears in the sample

  f_k is the frequency of a_k

- if we have n observations then:

  relative frequency = *frequency / n*

- *percentage relative frequency* = *relative frequency* × 100

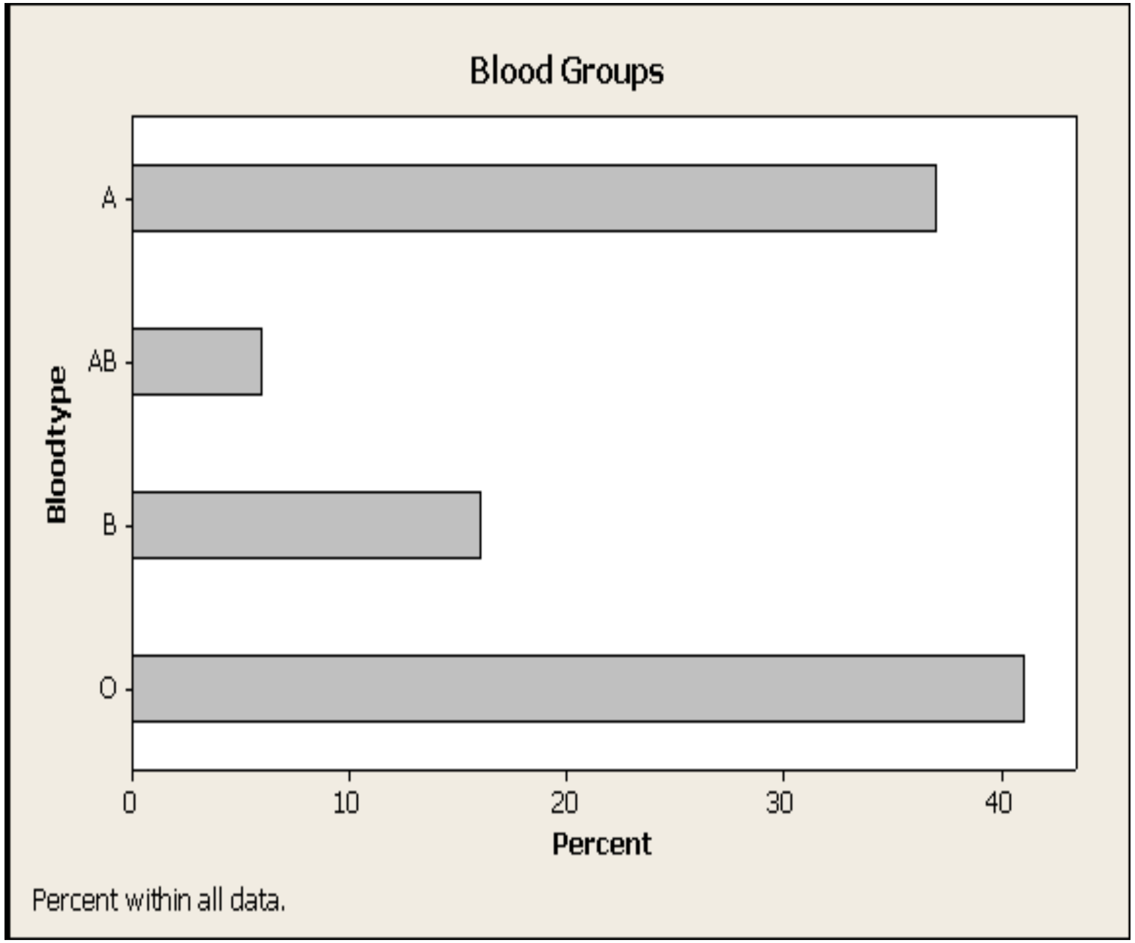
Frequency

sample of 572 patients   (*n = 572*)

Blood Type	Frequency	Relative Frequency	Percentage RF
A	210	0.37	37%
AB	35	0.06	6%
B	93	0.16	16%
O	234	0.41	41%
Totals	572	1.00	100%

sum of frequencies = n

Bar Chart


Blood Groups
A
AB
B
O
Bloodtype
0
10
20
30
40
Percent
Percent within all data.

Summarising data

Categorical (two variables):

- contingency table
- number of patients with blood type A who are female is 108

Blood Type	Sex		Totals
	male	female	
A	102	108	210
AB	12	23	35
B	46	47	93
O	120	114	234
Totals	280	292	572

Summarising data

Categorical (two variables):

- contingency table
- number of patients with blood type A who are female is 108

Blood Type	Sex		Totals	% Blood Type by sex	
	male	female		male	female
A	102	108	210	49%	51%
AB	12	23	35	34%	66%
B	46	47	93	50%	50%
O	120	114	234	51%	49%
Totals	280	292	572		

Bar Chart

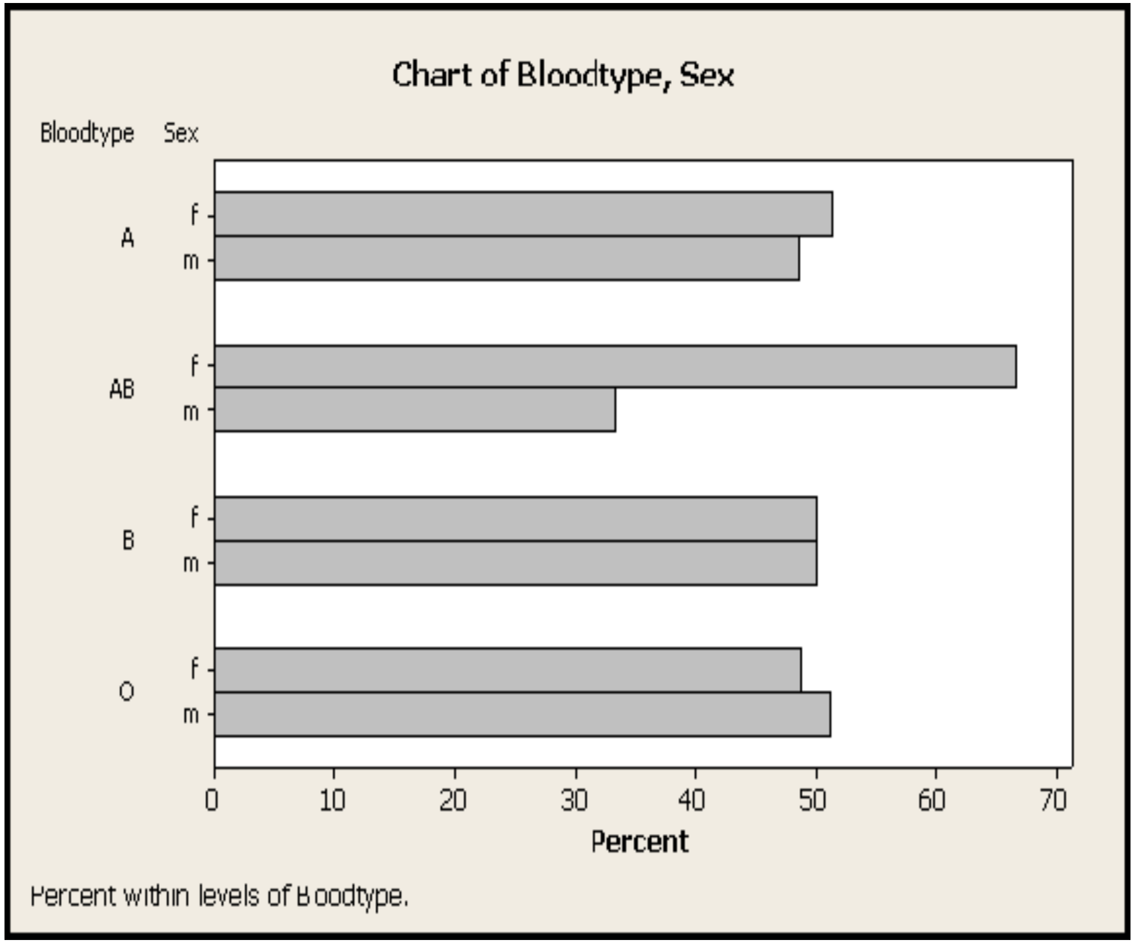


Chart of Bloodtype, Sex
Bloodtype
Sex
A
f
m
AB
f
m
B
f
m
O
f
m
0
10
20
30
40
50
60
70
Percent
Percent within levels of Boodtype.

Ordinal data

- X is an ordinal variable with values: $a_1,\ a_2,\ a_3,\ \ldots\ a_k,\ \ldots\ \ a_K$
- ‘ordinal’ means that:

$$a_1 \leq a_2 \leq a_3 \leq \ \ldots \leq \ a_k \leq \ \ldots \ \leq \ a_K$$

- cumulative frequency at level k:

  c_k = sum of frequencies of values less than or equal to a_k

$$\begin{aligned} c_k &= f_1 + f_2 + f_3 + \ \ldots \ + f_k \\ &= (f_1 + f_2 + f_3 + \ \ldots \ + f_{k-1}) + f_k \\ &= c_{k-1} + f_k \end{aligned}$$

- Can be applied to quantitative data as well ...

Cumulative frequencies

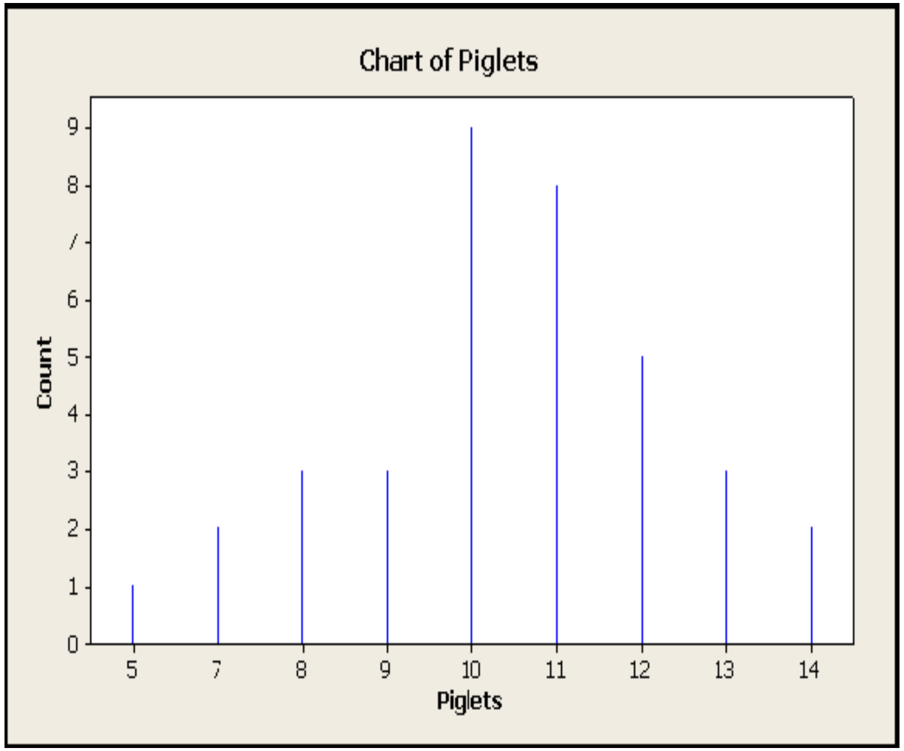
Number of piglets in a litter: (discrete data)

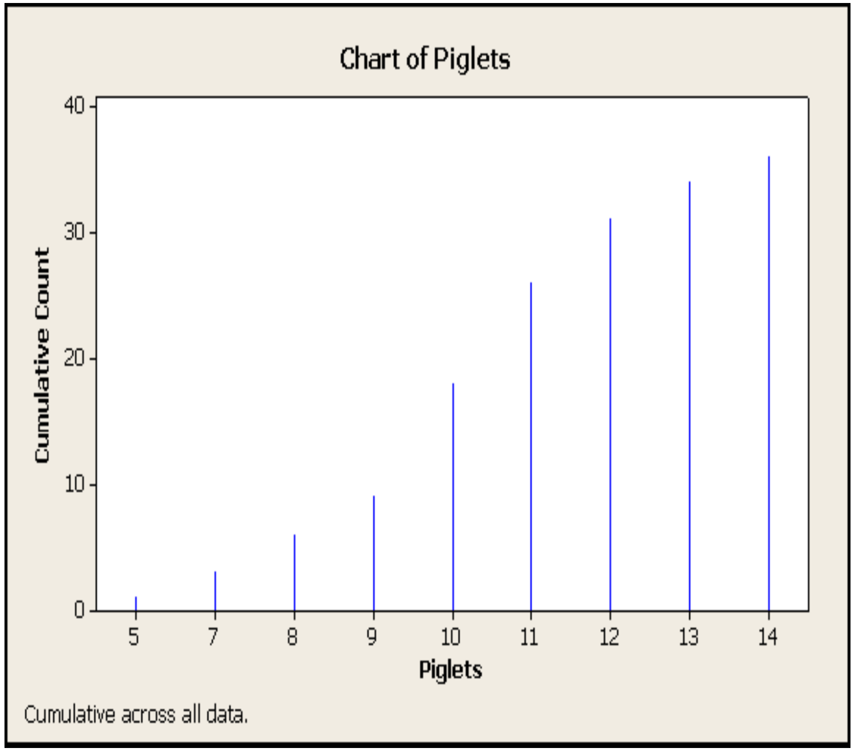
c1=f1=1,
c2=f1+f2=1,
c3=f1+f2+f3=3,
c4=f1+f2+f3+f4=6,
etc.

$$c_K = n$$

Litter size	Frequency=f	Cum. Freq =c
5	1	1
6	0	1
7	2	3
8	3	6
9	3	9
10	9	18
11	8	26
12	5	31
13	3	34
14	2	36
Total	36	

Plotting





frequency

cumulative frequency

Continuous data

- A way to obtain discrete numbers from continuous data:
  Divide range of observations into non-overlapping intervals (bins)
- Count number of observations in each bin

- Enzyme concentration data in 30 observations:

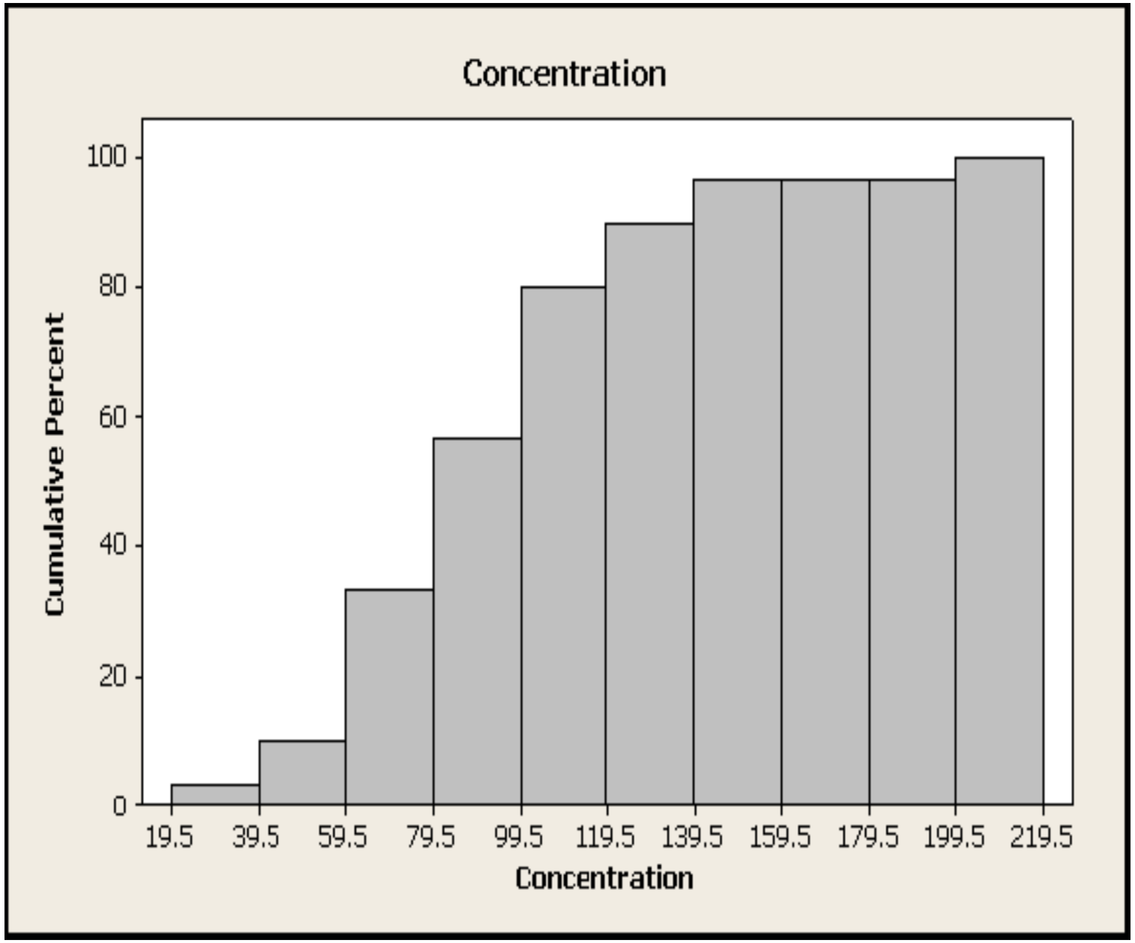
121	25	83	110	60	101
95	81	123	67	113	78
85	145	100	70	93	118
119	57	64	151	48	92
62	104	139	201	68	95

Range: 25 to 201 For example, you can use 10 bins of width 20:

Enzyme concentrations

Concentration	Freq.	Rel.Freq.	% Cum. Rel. Freq.
19.5 ≤ c < 39.5	1	0.033	3.3%
39.5 ≤ c < 59.5	2	0.067	10.0%
59.5 ≤ c < 79.5	7	0.233	33.3%
79.5 ≤ c < 99.5	7	0.233	56.6%
99.5 ≤ c < 119.5	7	0.233	79.9%
119.5 ≤ c < 139.5	3	0.100	89.9%
139.5 ≤ c < 159.5	2	0.067	96.6%
159.5 ≤ c < 179.5	0	0.000	96.6%
179.5 ≤ c < 199.5	0	0.000	96.6%
199.5 ≤ c < 219.5	1	0.033	100.0%
Totals	30	1.000	

Cumulative histogram



Ordinal data

- *X* is an ordinal variable with values: $a_1, \; a_2, \; a_3, \; \ldots \; a_k, \; \ldots \; a_K$
- 'ordinal' means that:

$$a_1 \le a_2 \le a_3 \le \ldots \le a_k \le \ldots \le a_K$$

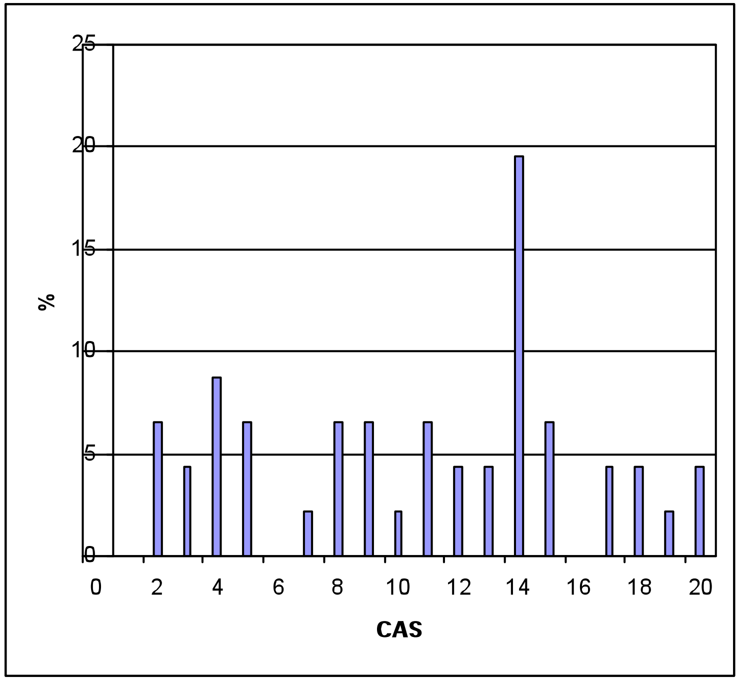
- cumulative frequency at level k:

  c_k = sum of frequencies of values less than or equal to a_k

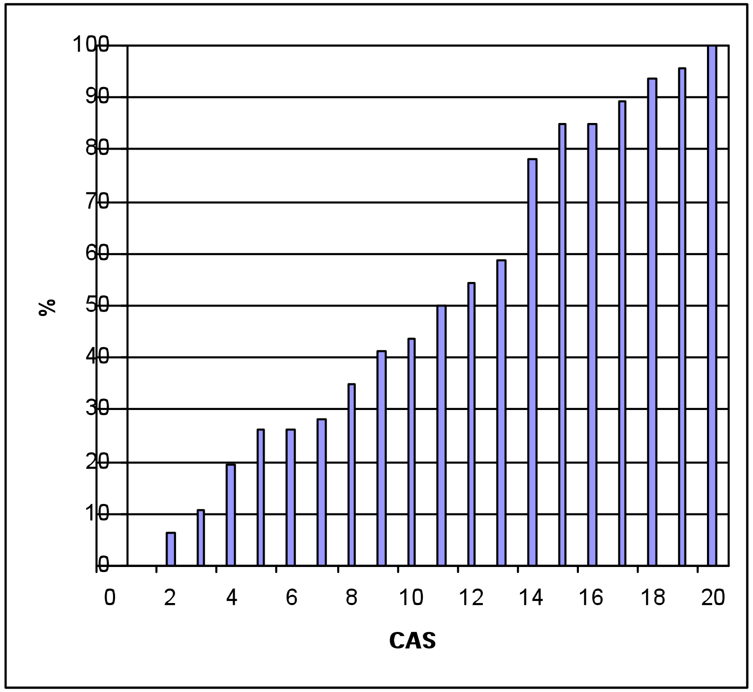
$$\begin{aligned} c_k &= f_1 + f_2 + f_3 + \ldots + f_k \\ &= (f_1 + f_2 + f_3 + \ldots + f_{k-1}) + f_k \\ &= c_{k-1} + f_k \end{aligned}$$

- also (%) cumulative relative frequency

CAS marks (last year)



% relative frequencies



% cumulative relative frequencies

A natural use of cumulative frequencies:

“What’s the percentage of students who failed?” ➔

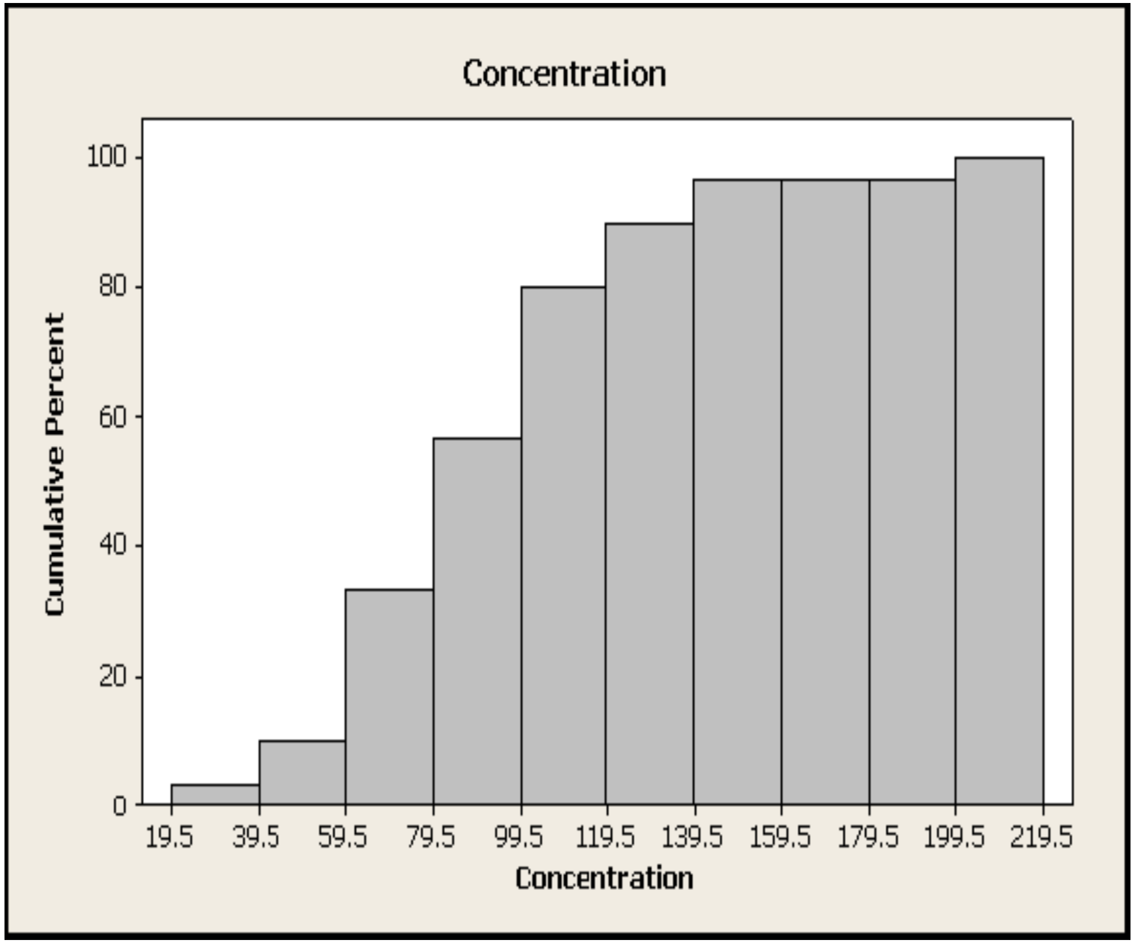
Look up the cumulative percentage at CAS 8 = c_9

$= f(a_1)$ =CAS 0 + $f(a_2)$ =CAS 1 + ... + $f(a_9)$ =CAS 8

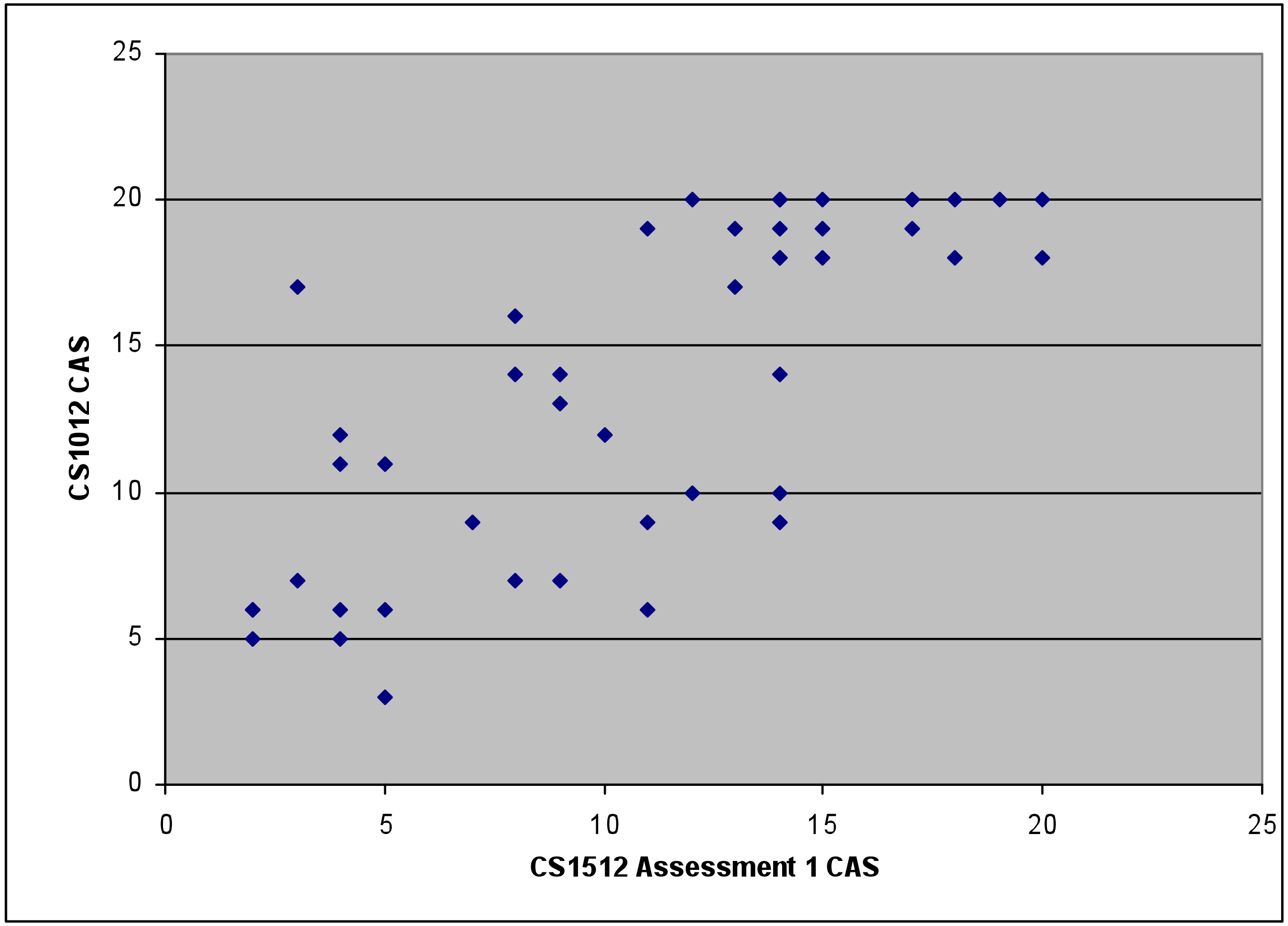
Enzyme concentrations

Concentration	Freq.	Rel.Freq.	% Cum. Rel. Freq.
19.5 ≤ c < 39.5	1	0.033	3.3%
39.5 ≤ c < 59.5	2	0.067	10.0%
59.5 ≤ c < 79.5	7	0.233	33.3%
79.5 ≤ c < 99.5	7	0.233	56.6%
99.5 ≤ c < 119.5	7	0.233	79.9%
119.5 ≤ c < 139.5	3	0.100	89.9%
139.5 ≤ c < 159.5	2	0.067	96.6%
159.5 ≤ c < 179.5	0	0.000	96.6%
179.5 ≤ c < 199.5	0	0.000	96.6%
199.5 ≤ c < 219.5	1	0.033	100.0%
Totals	30	1.000	

Cumulative histogram



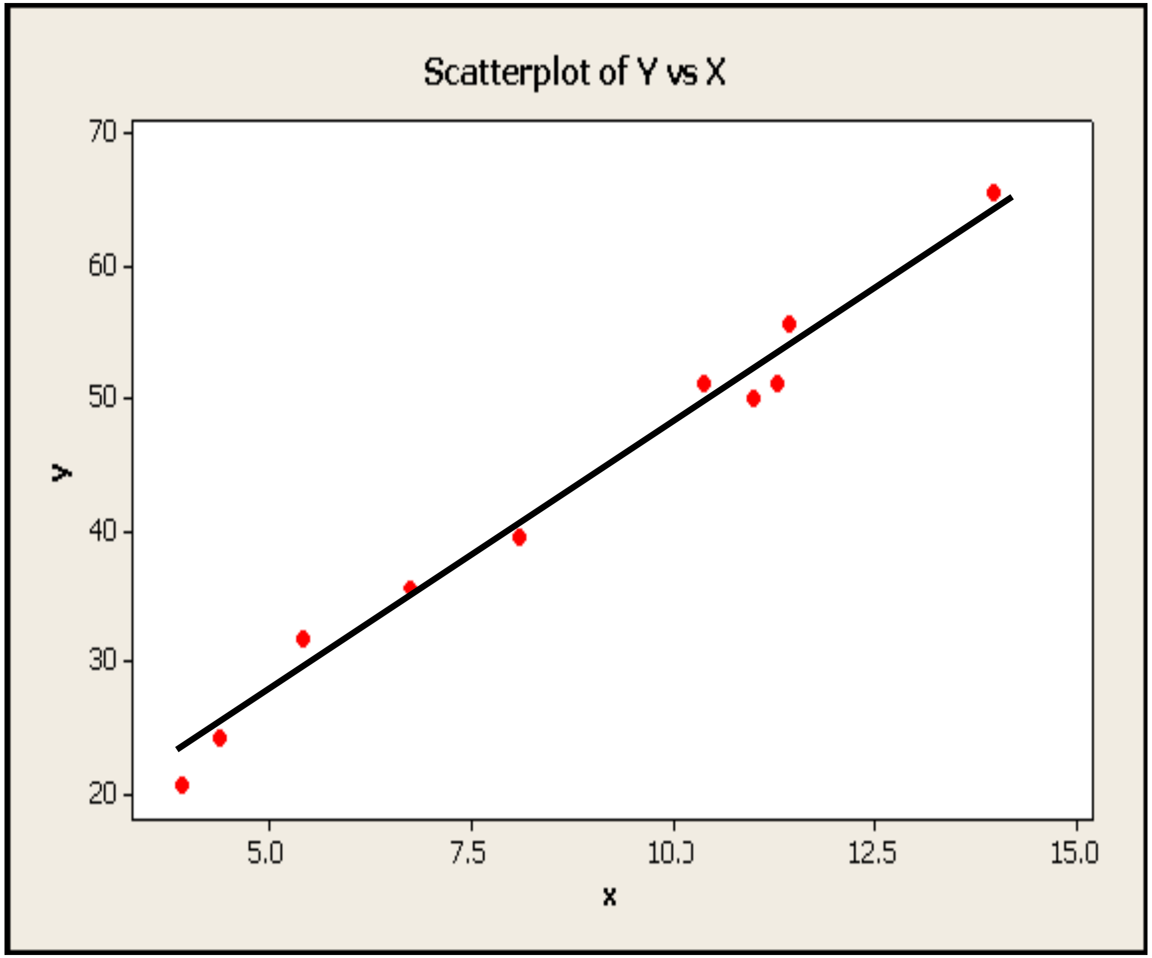
Discrete two variable data



What would you see if students tended to get the same mark for the two courses?

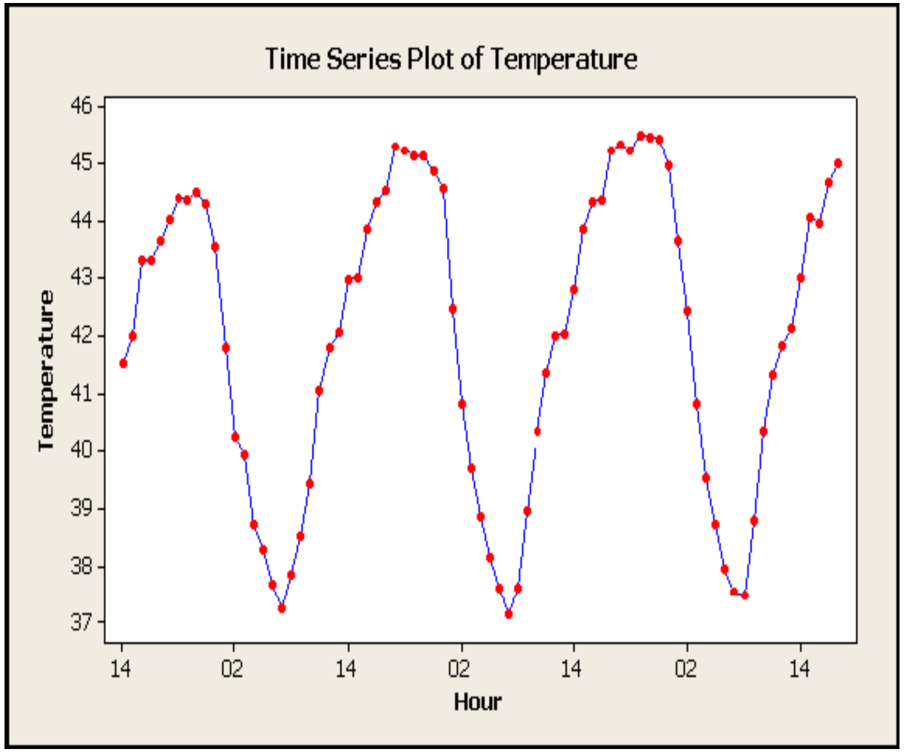
Continuous two variable data

X	Y
4.37	24.19
8.10	39.57
11.45	55.53
10.40	51.16
3.89	20.66
11.30	51.04
11.00	49.89
6.74	35.50
5.41	31.53
13.97	65.51



Time Series

- Time and space are fundamental (especially time)
- Time series: variation of a particular variable with time



Summarising data by numerical means

Further summarisation (beyond frequencies)
No inference yet!

Measures of location (Where is the middle?)

- Mean
- Median
- Mode

Mean

$$\text{Sample Mean } (\overline{X}) = \frac{\text{sum of observed values of X}}{\text{number of observed values}}$$

$$= \frac{\sum x}{n}$$

Mean

$$\text{Sample Mean } (\bar{X}) = \frac{\text{sum of observed values of X}}{\text{number of observed values}}$$

$$= \frac{\sum x}{n}$$

use only for quantitative data

Sigma

Sum of n observations

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_i + \ldots + x_{n-1} + x_n$$

If it is clear that the sum is from 1 to n then we can use a shortcut:

$$\sum x = x_1 + x_2 + \ldots + x_i + \ldots + x_{n-1} + x_n$$

Sigma

Sum of n observations

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \dots + x_i + \dots + x_{n-1} + x_n$$

If it is clear that the sum is from 1 to *n* then we can use a shortcut:

$$\sum x = x_1 + x_2 + \dots + x_i + \dots + x_{n-1} + x_n$$

Sum of squares (using a similar shortcut)

$$\sum x^2 = x_1^2 + x_2^2 + \dots + x_i^2 + \dots + x_{n-1}^2 + x_n^2$$

Mean from frequencies

What if your data take the form of frequencies: a_1 occurs f_1 times, a_2 occurs f_2 times, etc.?

Group together those x's which have value a_1, those with value a_2, ...

$$\begin{array}{lll}
\sum x = & x_{..} + x_{..} + x_{..} \ \ldots + & x\text{'s which have value } a_1 \text{ - there are } f_1 \text{ of them} \\
 & x_{..} + x_{..} \ \ldots + & x\text{'s which have value } a_2 \text{ - there are } f_2 \text{ of them} \\
 & \ldots & \\
 & x_{..} + x_{..} & x\text{'s which have value } a_K \text{ - there are } f_K \text{ of them}
\end{array}$$

$$= f_1 * a_1 + f_2 * a_2 + \ldots + f_k * a_k + \ldots + f_K * a_K$$

$$= \sum_{k=1}^{K} f_k * a_k$$

Mean

Litter size a_k	Frequency f_k	Cum. Freq
5	1	1
6	0	1
7	2	3
8	3	6
9	3	9
10	9	18
11	8	26
12	5	31
13	3	34
14	2	36
Total	36	

$$\sum x = \sum_{k=1}^{K} f_k * a_k$$

$$\begin{aligned} &= 1*5 \quad + 0*6 \quad + 2*7 + 3*8 \\ &\quad 3*9 \quad + 9*10 + 8*11 \\ &\quad 5*12 + 3*13 + 2*14 \\ &= 375 \end{aligned}$$

$$\bar{X} = 375 / 36$$

$$= 10.42$$

Another kind of middle: the Median

Sample median of X = middle value when n sample observations
are ranked in increasing order
= the $((n + 1)/2)^{\text{th}}$ value

Equally many values on both sides

n odd: values: 183, 185, 184
rank order: 183, 184, 185
median: 184

n odd: values: 183, 200, 184
rank order: 183, 184, 200
median: 184 *Median doesn't care about outliers!*

Another kind of middle: the Median

Sample median of X = middle value when n sample observations are ranked in increasing order

= the ((n + 1)/2)th value

n even: values: 183,200,184,185
rank order: 183,184,185,200
median: (184+185)/2 = 184.5

(When n is even, there is no (n+1)/2)th value: in this case, the median is the mean of the two values "surrounding" the nonexistent (n+1)/2)th value. In our example, that's (184+185)/2))

CS1512

Another kind of middle: the Median

Sample median of X = middle value when n sample observations are ranked in increasing order

= the ((n + 1)/2)th value

n odd: values: 183, 163, 152, 157 and 157
rank order: 152, 157, 157, 163, 183
median:

n even: values: 165, 173, 180, 164
rank order: 164, 165, 175, 180
median:

UNIVERSITY OF ABERDEEN

Another kind of middle: the Median

Sample median of X = middle value when n sample observations are ranked in increasing order

= the $((n + 1)/2)^{th}$ value

n odd:	values:	183, 163, 152, 157 and 157
	rank order:	152, 157, 157, 163, 183
	median:	157
n even:	values:	165, 175, 180, 164
	rank order:	164, 165, 175, 180
	median:	(165 + 175)/2 = 170

Median

Litter size	Frequency	Cum. Freq
5	1	1
6	0	1
7	2	3
8	3	6
9	3	9
10	9	18
11	8	26
12	5	31
13	3	34
14	2	36
Total	36	

Median = 10.5

Mode

Sample mode = value with highest frequency (may not be unique)

Litter size	Frequency	Cum. Freq
5	1	1
6	0	1
7	2	3
8	3	6
9	3	9
10	9	18
11	8	26
12	5	31
13	3	34
14	2	36

Mode = ?

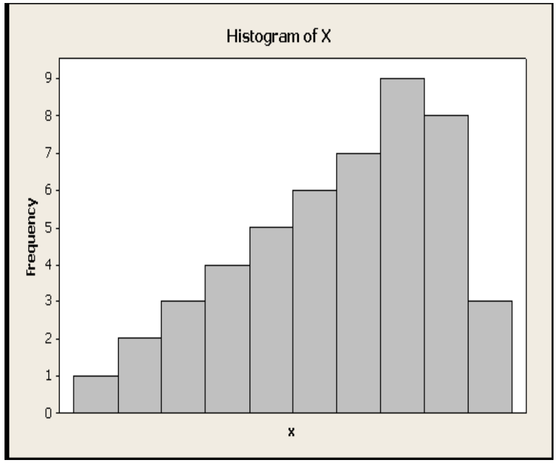
Mode

Sample mode = value with highest frequency (may not be unique)

Litter size	Frequency	Cum. Freq
5	1	1
6	0	1
7	2	3
8	3	6
9	3	9
10	9	18
11	8	26
12	5	31
13	3	34
14	2	36

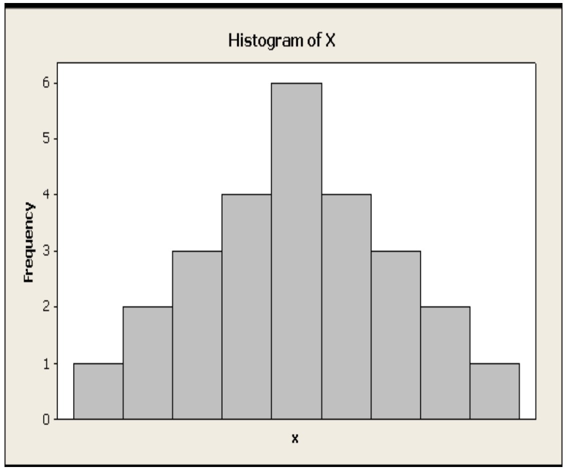
Mode = 10

Skew in histograms



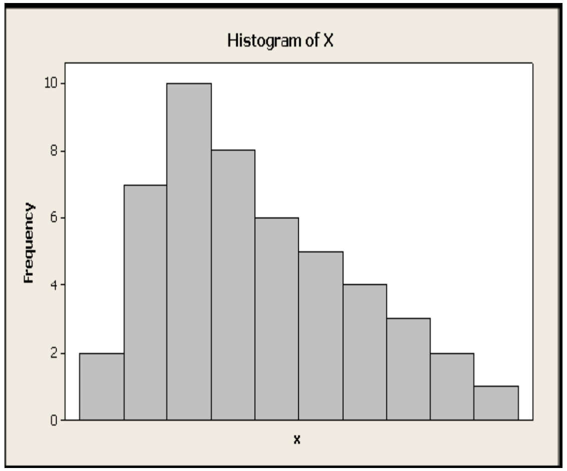
left skewed

mean < mode



symmetric

mean ≈ mode



right skewed

mean > mode

How much variation is there in my data?

We’ve seen various ways of designating the `middle value` *(mean, median, mode)*

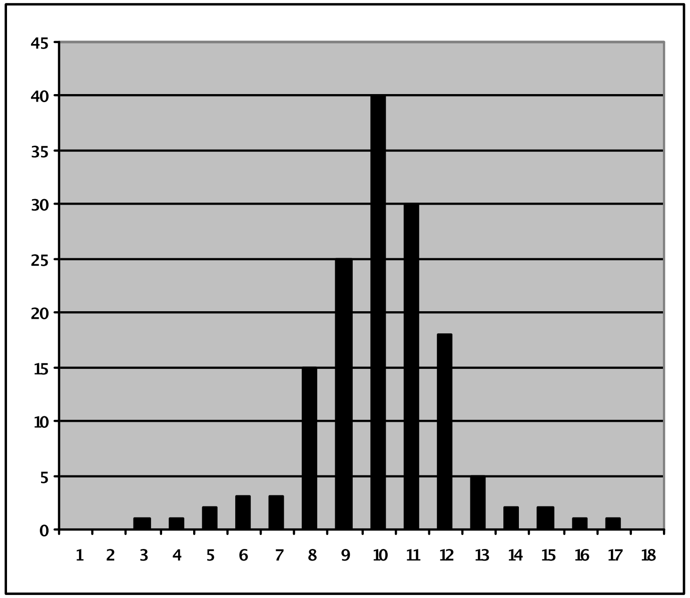
Sometimes most values are close to the mean, sometimes they are not.

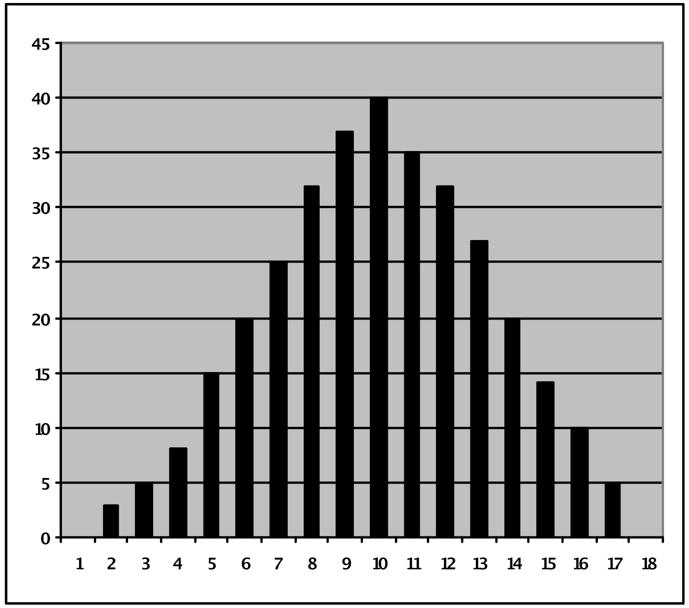
How can we quantify how close the values are (on average) to the mean? (We’re looking for a measure of “spread”)

First we introduce variance, then the measure most often used, called *Standard Deviation*

Variance

Measure of spread: variance





Variance

sample variance = v, also called s^2 (you will see why)

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)$$

(Don’t use when n=1. In this case, v=0.)

sample standard deviation = s = $\sqrt{\text{variance}}$

Variance

sample variance = v, also called s^2

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)$$

Why $(..-..)^2$? *That's because we're interested in the absolute distances to the mean. (If we summated positive and negative distances, the sum would always be 0.) When standard deviation takes the root of v, you can think of that as correcting the increase in values caused by the formula for v.*

Why divide by n-1? We want the average distance, so we need to take the number n of values into account. (n-1 gives more intuitive values than n, particularly when n is small)

A trick for calculating Variance

(equation stated without proof)

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)$$

$$s^2 = \frac{1}{n-1}\left(\sum x^2 - \frac{(\sum x)^2}{n}\right).$$

Variance and standard deviation

Litter size a_k	Frequency f_k	Cum. Freq
5	1	1
6	0	1
7	2	3
8	3	6
9	3	9
10	9	18
11	8	26
12	5	31
13	3	34
14	2	36
Total	36	

$$\sum x^2 = \sum_{k=1}^{K} f_k * a_k^2$$

$$\begin{aligned} &= 1*25 + 2*49 + 3*64 \\ &\quad 3*81 + 9*100 + 8*121 \\ &\quad 5*144 + 3*169 + 2*196 \\ &= 4145 \end{aligned}$$

$$\sum x = 375$$

$$(\sum x)^2 / n = 375*375 / 36$$

$$= 3906$$

$$s^2 = (4145-3906) / (36-1)$$

$$= 6.83$$

$$s = 2.6$$

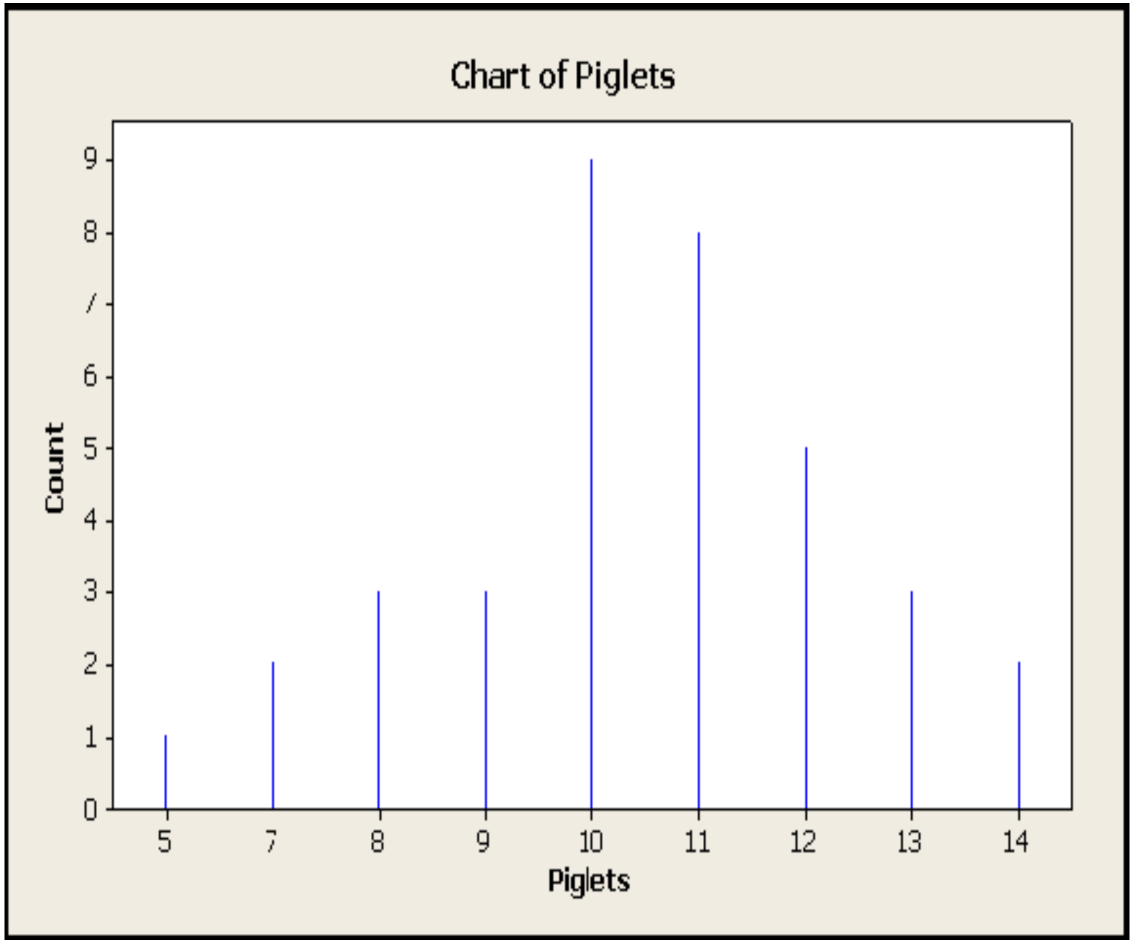
Variance/standard deviation

NB: In practice, these are seldom calculated by hand. Software packages like Excel perform these (and much harder) calculations automatically – But it’s useful to do it yourself a few times.

Question to think about at home

What happens with SD if you double the values of all variables?
(Does SD stay the same?)

Piglets



Mean = 10.42

Median = 10.5

Mode = 10

Std. devn. = 2.6

Standard deviation gives you a "global" perspective on spread (i.e. how much spread there is in the sample as a whole)

Sometimes what's most striking about your data is not how much spread there is, but that the data are very skew

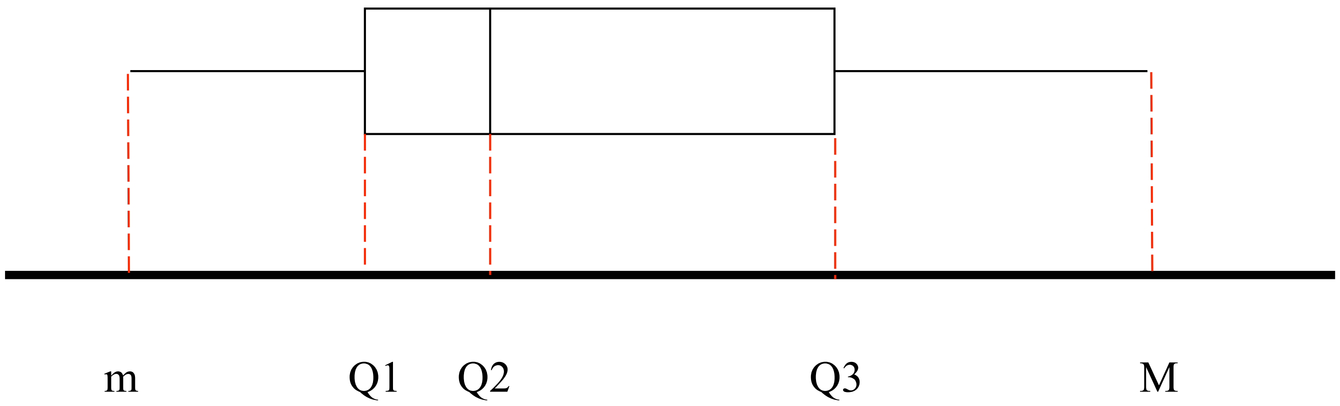
In those cases, quartiles can give insight

Quartiles and Range

Median: value such that 50% of observations are **below** (**above**) it (Q2).
Lower quartile: value such that 25% of observations are **below** it (Q1).
Upper quartile: value such that 25% of observations are **above** it (Q3).

Range: the **minimum** (m) and **maximum (M)** observations.

Box and Whisker plot:



Quartiles and Range

Defined more precisely in the same way as median:

- Lower quartile = the $((n+1)/4)^{th}$ value
- Upper quartile = the $(3(n+1)/4)^{th}$ value

See D.G. Rees, p.40. Example: five people’s heights:
{183cm,163cm,152cm,157cm,157cm}.

Quartiles and Range

Defined more precisely in the same way as median:

- Lower quartile = the ((n+1)/4)th value
- Upper quartile = the (3(n+1)/4)th value

See D.G. Rees, p.40. Example: five people's heights: {183cm,163cm,152cm,157cm,157cm}. Arranged in rank order: {152cm,157cm,157cm,163cm,183cm}. Since n=5,
LQ=the ((5+1)/4)th value

Quartiles and Range

Defined more precisely in the same way as median:

- Lower quartile = the ((n+1)/4)th value
- Upper quartile = the (3(n+1)/4)th value

See D.G. Rees, p.40. Example: five people's heights: {183cm,163cm,152cm,157cm,157cm}. Arranged in rank order: {152cm,157cm,157cm,163cm,183cm}. Since n=5,
LQ=the ((5+1)/4)th value=the 1.5th value=
the mid point between 152 and 157=
(152+157)/2=309/2=154.5

Linear Regression

Recall the situation where you try to relate two variables, such as
x=each student's score on the CS1012 exam
y=each student's score on the CS1512 exam

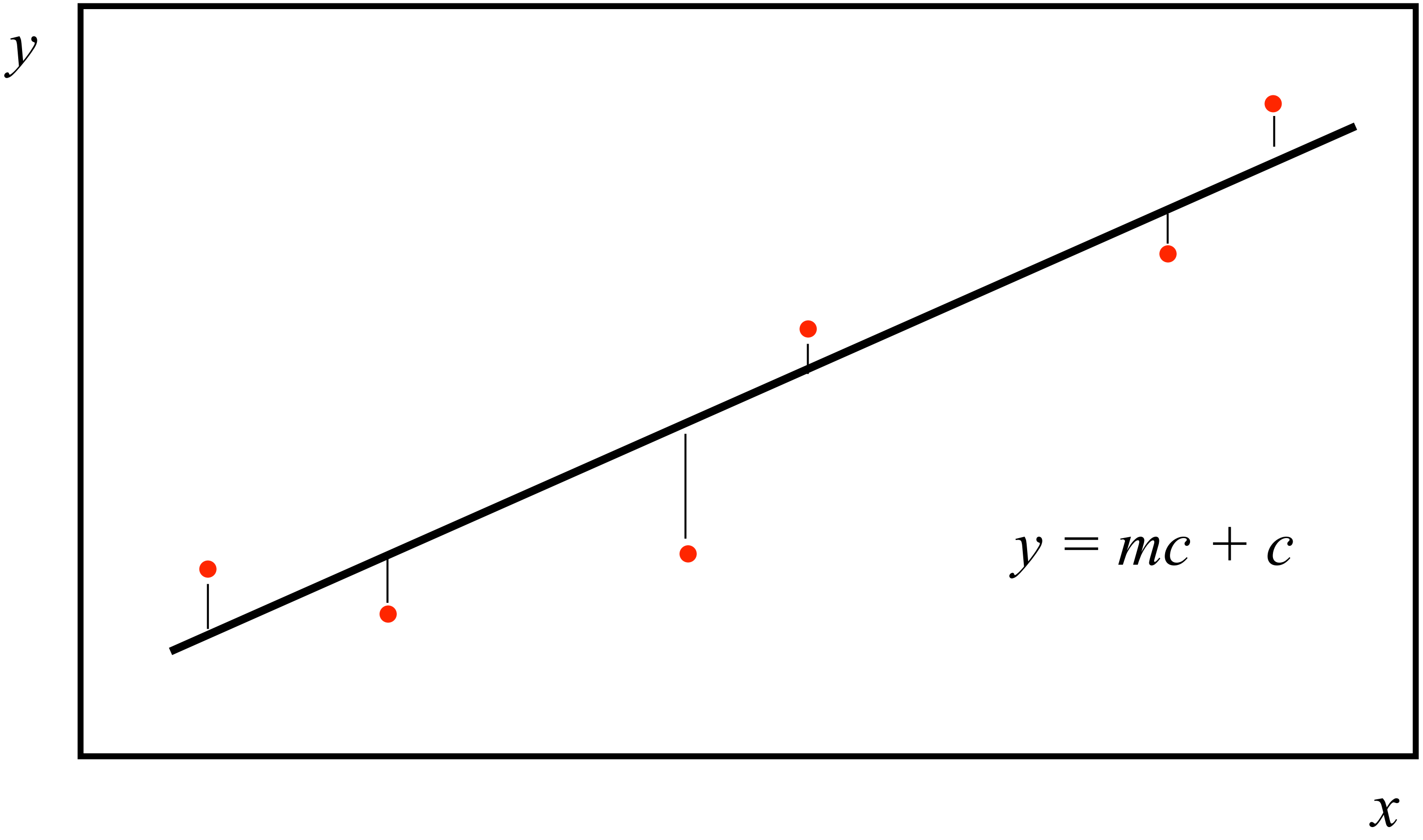
We have seen: If these are positively related (linearly), then their graph will approximate a straight line

The simplest case occurs when each students has *the same* score for both exams, in which case the graph will coincide with the diagonal x=y.

If the graph only approximates a straight line, then how closely does it approximate the line?

Linear Regression

Calculate *m* and *c* so that $\sum$(distance of point from line)^2 is minimised



Linear Regression

Observe that *Linear Regression* is based on the same idea as the notions of *Variance* and *Standard Deviation*: summation of squared distances (from something)

Structured sample spaces

Sometimes you don’t want to throw all your data on one big heap, for example because they represent observations concerning different points in time

Does this make it meaningless to talk about the sample mean?

Structured sample spaces

Sometimes you don’t want to throw all your data on one big heap, for example because they represent observations concerning different points in time

Does this make it meaningless to talk about the sample mean?

No, you may still want to know the mean as calculated over the set of all time points.

Structured sample spaces

Sometimes you don’t want to throw all your data on one big heap, for example because they represent observations concerning different points in time

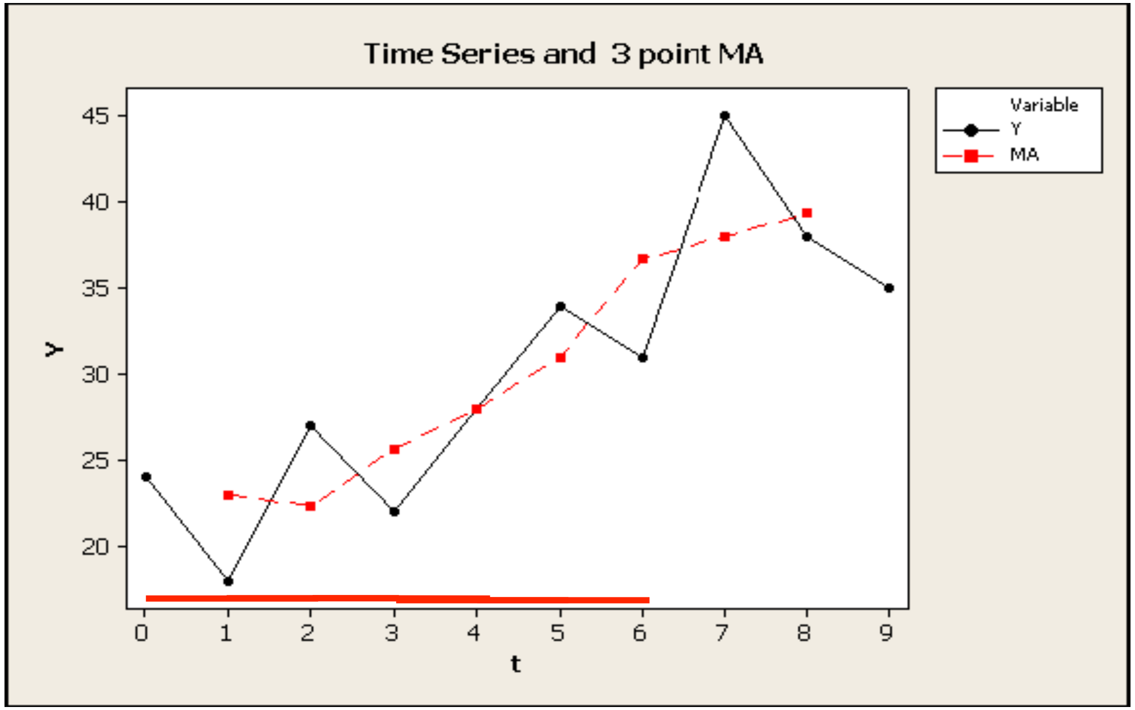
Does this make it meaningless to talk about the sample mean?

No, you may still want to know the mean as calculated over the set of all time points.

Or, you may want to know the mean for some smaller collections of time points. Example: the Moving Average:

Time Series - Moving Average

Time	Y	3 point MA
0	24	*
1	18	23.0000
2	27	22.3333
3	22	25.6667
4	28	28.0000
5	34	31.0000
6	31	36.6667
7	45	38.0000
8	38	39.3333
9	35	*



- smoothing function
- can compute median, max, min, std. devn, etc. in window

Next: Probability