CX2100 Probability & Statistics

Part 1 Instructor: A/P Lau Chiew Tong
Room: N4-2B-58; Email: asctlau@ntu.edu.sg

Part 2 Instructor: A/P Adams Kong
Room: N4-2A-17; Email: adamskong@ntu.edu.sg

~ 24 Sessions of lectures

~ 12 Sessions of tutorials

Assessments (to be confirmed):

- 4 Quizzes – 25% each
- No Final exam

Part 1 (Week 1-7):

Ch1 Introduction to Statistics

Ch2 Presenting Data

Ch3 Summarizing Distributions

Ch4 Bivariate Data

Ch5 Probability Theory

Ch6 Probability Distribution

- Random Variables
- Discrete Distribution
- Continuous Distribution

Part 2 (Week 8-13):

To be taught by Prof Adams Kong

- TEL supporting Materials for Ch1-4 (short video clips and notes) are available in NTULearn. Students are advised to watch the video clips before attending the lectures for Part 1 (Ch1-4).

- Lectures (Week 1 to 7)
  - Discussion & worked examples on topics covered in the TEL materials
  - Additional topics not covered in TEL materials (Ch 5 and 6)

- Recess week (Self-study non-exam)
  - Use of R programming for analysis and presentation of statistical data (Practice materials will be available in NTULearn)

- Tutorial – One session per week, to begin in Week 3

All Course Materials are available in NTULearn.

Prof Kong will take over the teaching from Wk 8.

Students participation during lectures using Student Response System - Wooclap

A QR code or URL will be presented for students to join Wooclap.

Scan the QR code with your mobile phone and access the web page. Account registration is not required.

Let's try out the polling session by answering this question using your mobile phone:

Ch 1. Introduction to Statistics

- Descriptive & Inferential Statistics
- Types of Variables
- Percentiles
- Types of Measurement Scale
- Distributions
- Linear Transformations

Statistics involve gathering, organizing, analyzing, interpreting and presenting data.

Statistics are being used everywhere:

- Number of students enrolled in this course
- Index measuring stock market
- Singapore household income
- Live data of new Covid-19 cases
- Opinion poll, benchmarks poll, tracking polls, etc.

Example: A toothpaste manufacturer claims that more than 80% of Dentists recommend a particular brand of toothpaste. This was based on surveys of dentists which allow selection of one or more brands.

Is the claim misleading?

Yes.

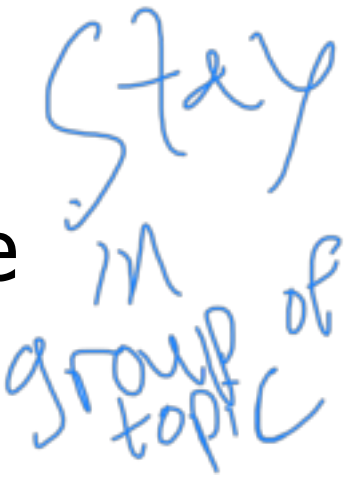
Because it may be understood that 80% of dentists recommend this brand over the others. It should be noted that other brands were also recommended and may be as much as that particular brand.

Why Statistics are important?

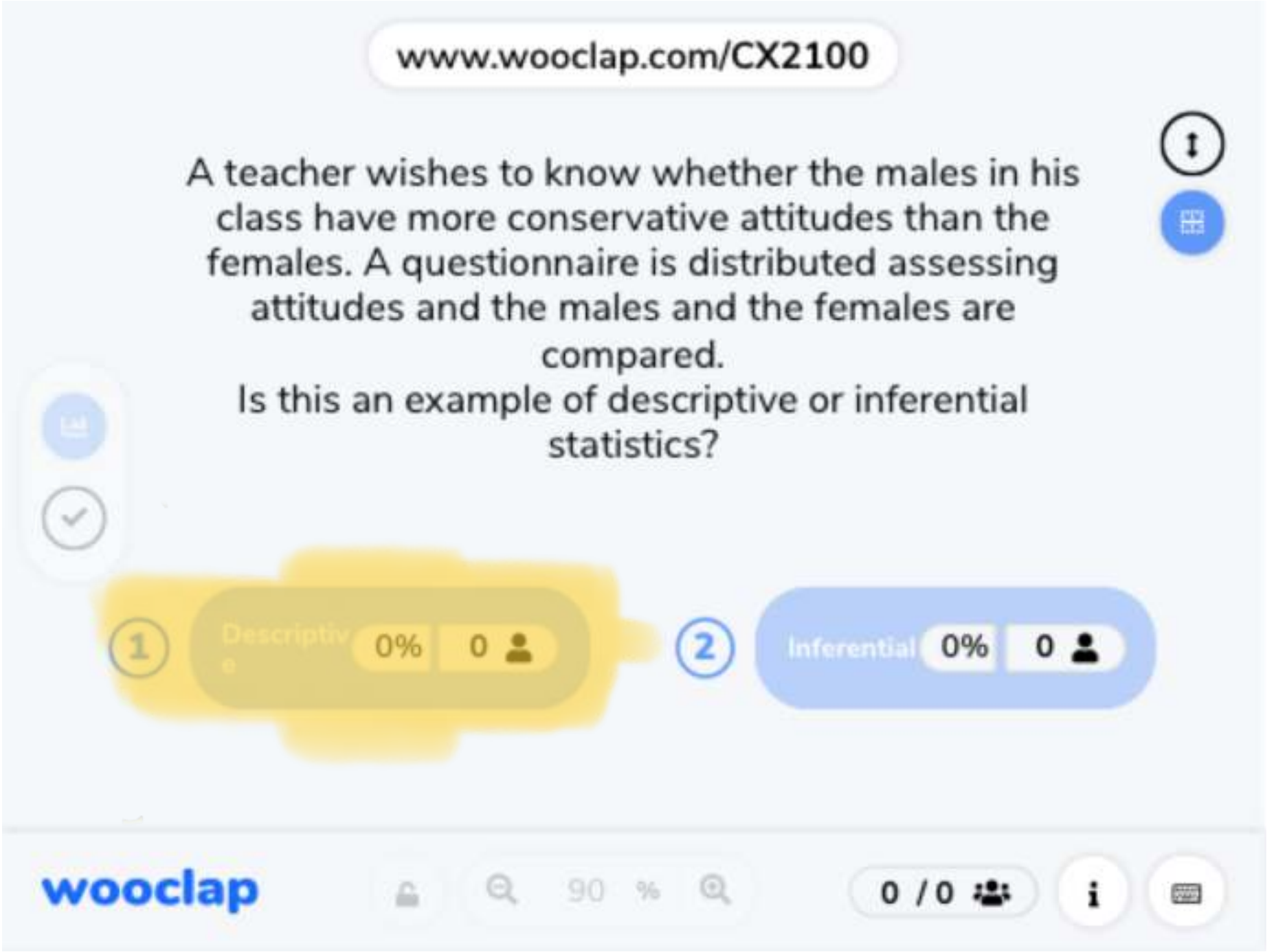
- Predicting the spreading of diseases
- Weather forecasting based on statistics
- Provide informed choice on investment decision
- AI or machine learning based on past statistical data

- Descriptive & Inferential Statistics

Descriptive statistics – summarize and describe important features of the data collected. Does not generalize beyond the data collected.



Inferential statistics - collection of sample to draw inferences about the population, i.e. formal guesses of statistical parameters about the population by looking at the samples.


www.wooclap.com/CX2100
A teacher wishes to know whether the males in his class have more conservative attitudes than the females. A questionnaire is distributed assessing attitudes and the males and the females are compared.
Is this an example of descriptive or inferential statistics?
1
0%
0
2
Inferential
0%
0
wooclap
90 %
0 / 0

A cognitive psychologist is interested in comparing two ways of presenting stimuli on subsequent memory. Twelve subjects are presented with each method and a memory test is given.

What would be the roles of descriptive and inferential statistics in the analysis of these data?

Descriptive statistics – we describe and analyze the data from the sample.

Inferential statistics – we use the data from the sample to generalize to a larger population of people.

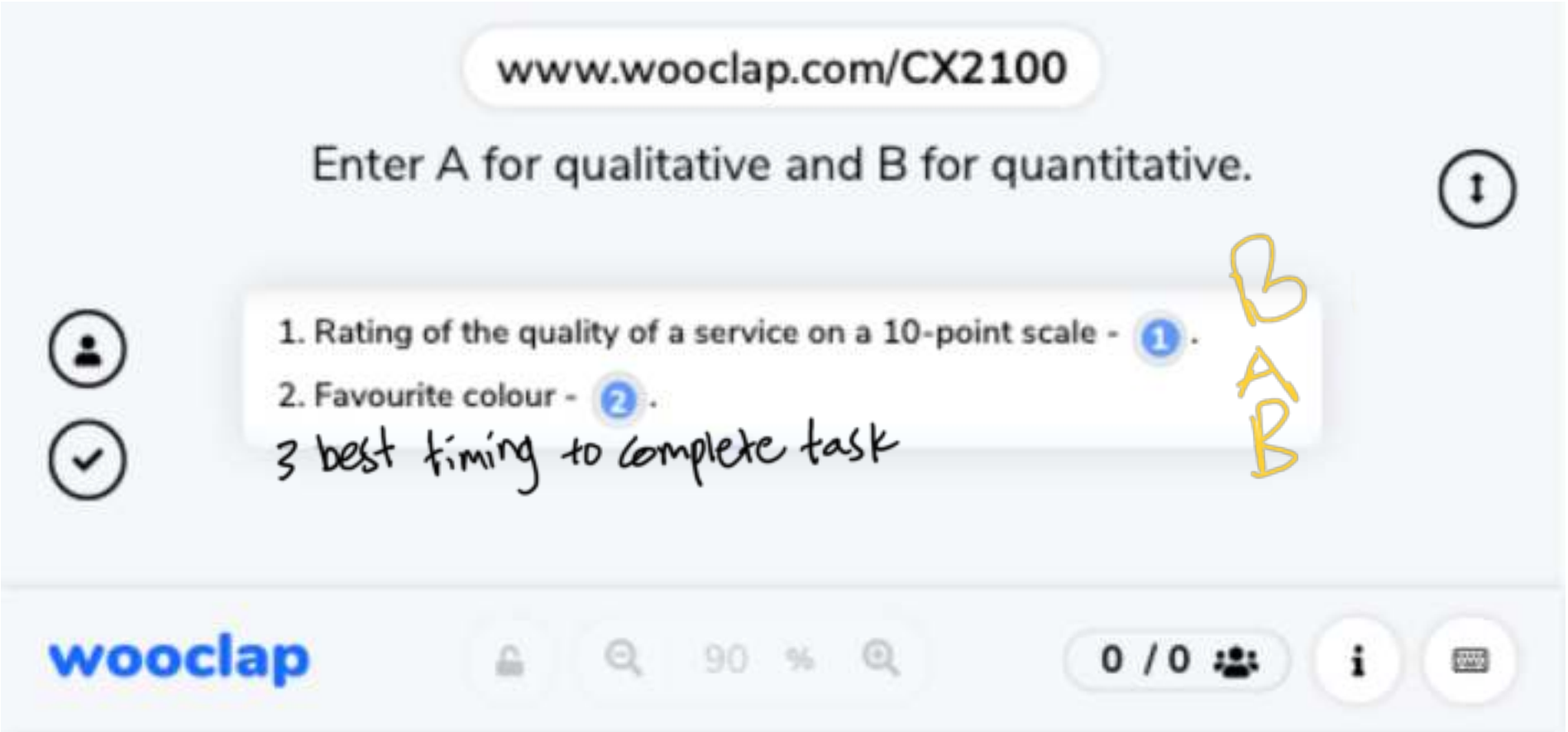
- Types of Variables

In statistic, we can broadly group variable into two categories: Qualitative and Quantitative.

express in name, etc

express in number

Categorize the following variables as being qualitative or quantitative.



- Percentile

In certain experiments, it is more meaningful to compare the outcomes obtained.

Eg: if you know that your quiz marks is 65 out of 100, you may not know how well you have done compared to others in your class.

A **percentile** is a comparison score between a particular score and the scores of the rest of a group.

Percentile - calculation

Calculation of P^{th} Percentile for a set of N data:

Data arranged in the order of magnitude

1. Compute the rank $R = \frac{P}{100} \times (N + 1)$
2. Let I_R = Integer part of R and F_R = Fractional part of R
3. P^{th} Percentile = Data at rank I_R +

   (Data at rank (I_R+1) – Data at rank I_R) x F_R

Eg: Given data: [3, 5, 7, 8, 9, 11, 13, 15], compute the 25^{th} and 75^{th} percentile.

$25^{th} = \frac{1}{4}(9) = 2.25$

$5 + (7-5)\,0.25 = 5.5$

$75^{th} = \frac{75}{100}(9) = \frac{27}{4} = 6.75$

$11 + (13-11)\,0.75$

$11 + 1.5 = 12.5$

Given a data set of 20 values:

1	2	3	4	5	5	6	6	6	7
7	8	8	8	9	9	9	10	10	10

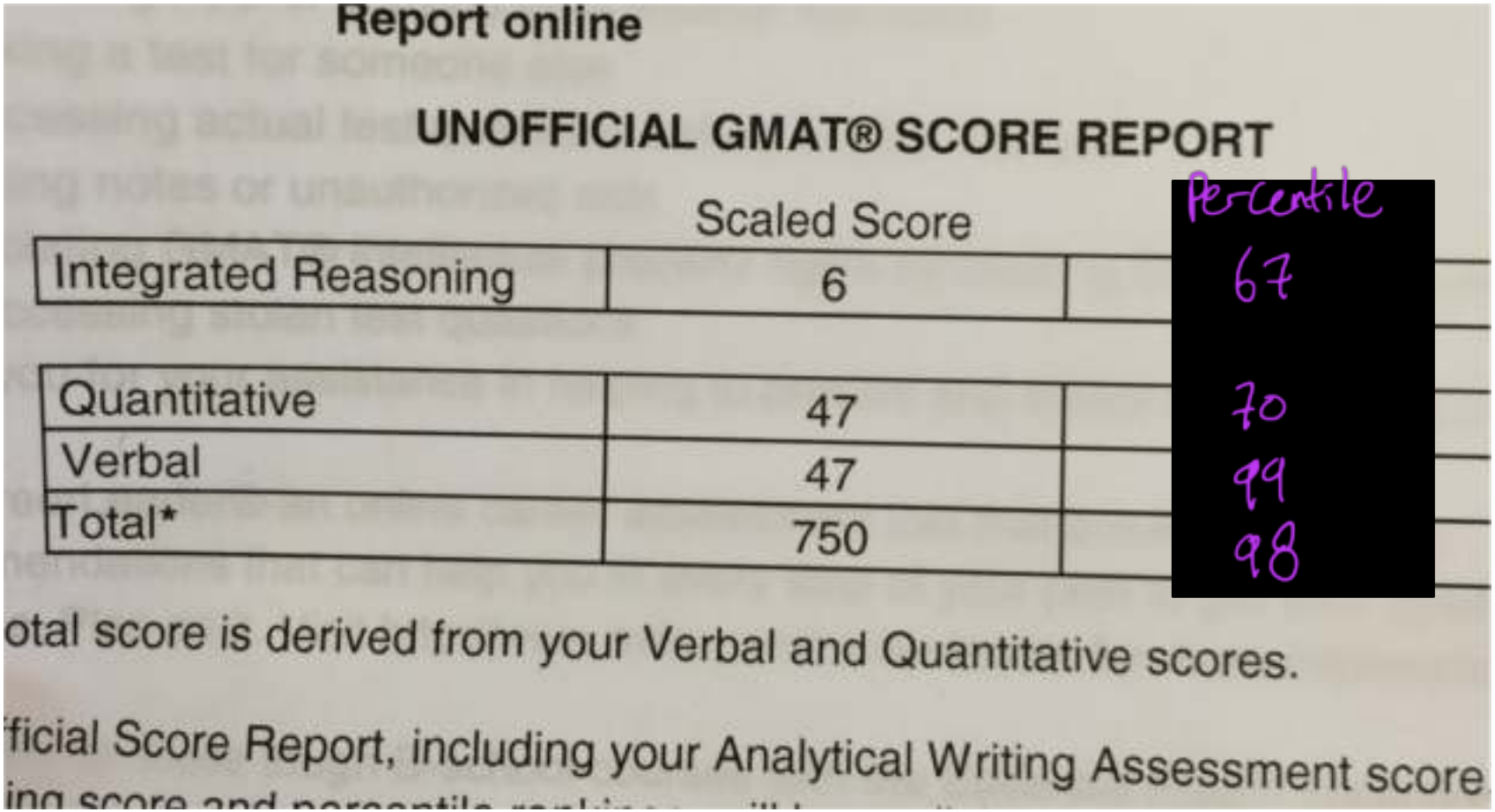
need to make sure the order then calculate

Given a data set of 20 values:

1	2	3	4	5	5	10	10	10	6
6	9	9	9	7	7	7	8	8	8

Practical example:

Graduate Management Admissions Test - a standardized test for application to graduate-level business programs.



Report online

UNOFFICIAL GMAT® SCORE REPORT

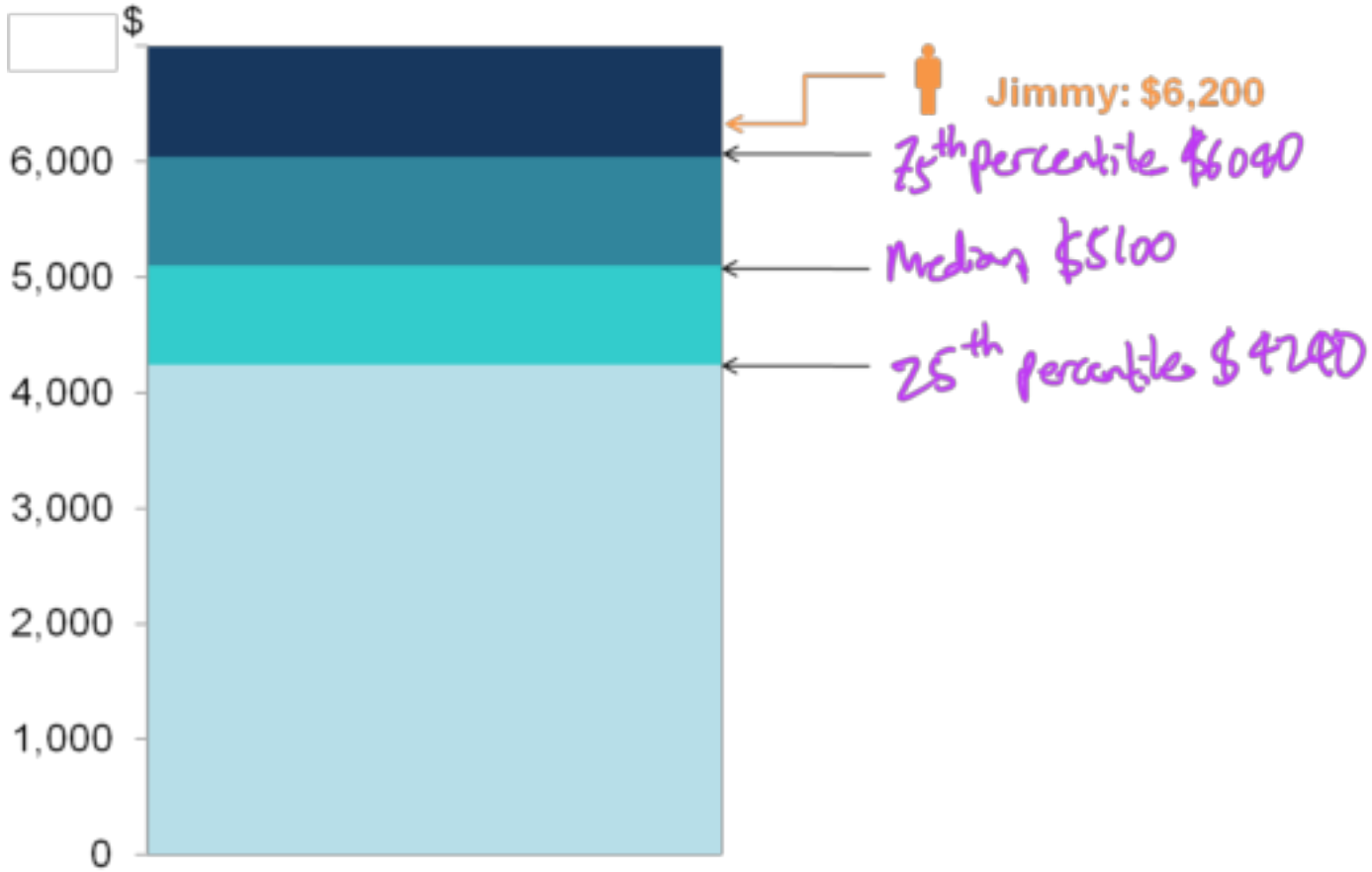
	Scaled Score	Percentile
Integrated Reasoning	6	67
Quantitative	47	70
Verbal	47	99
Total*	750	98

otal score is derived from your Verbal and Quantitative scores.

ficial Score Report, including your Analytical Writing Assessment score

Example: Percentile for benchmarking.

Suppose Jimmy earns a monthly wage of $6,200. He would like to compare his wages with those working in the same occupation and industry.



https://stats.mom.gov.sg/SL/Pages/Percentiles-Uses.aspx

Types of Measurement Scales

Four basic levels of measurement scales:

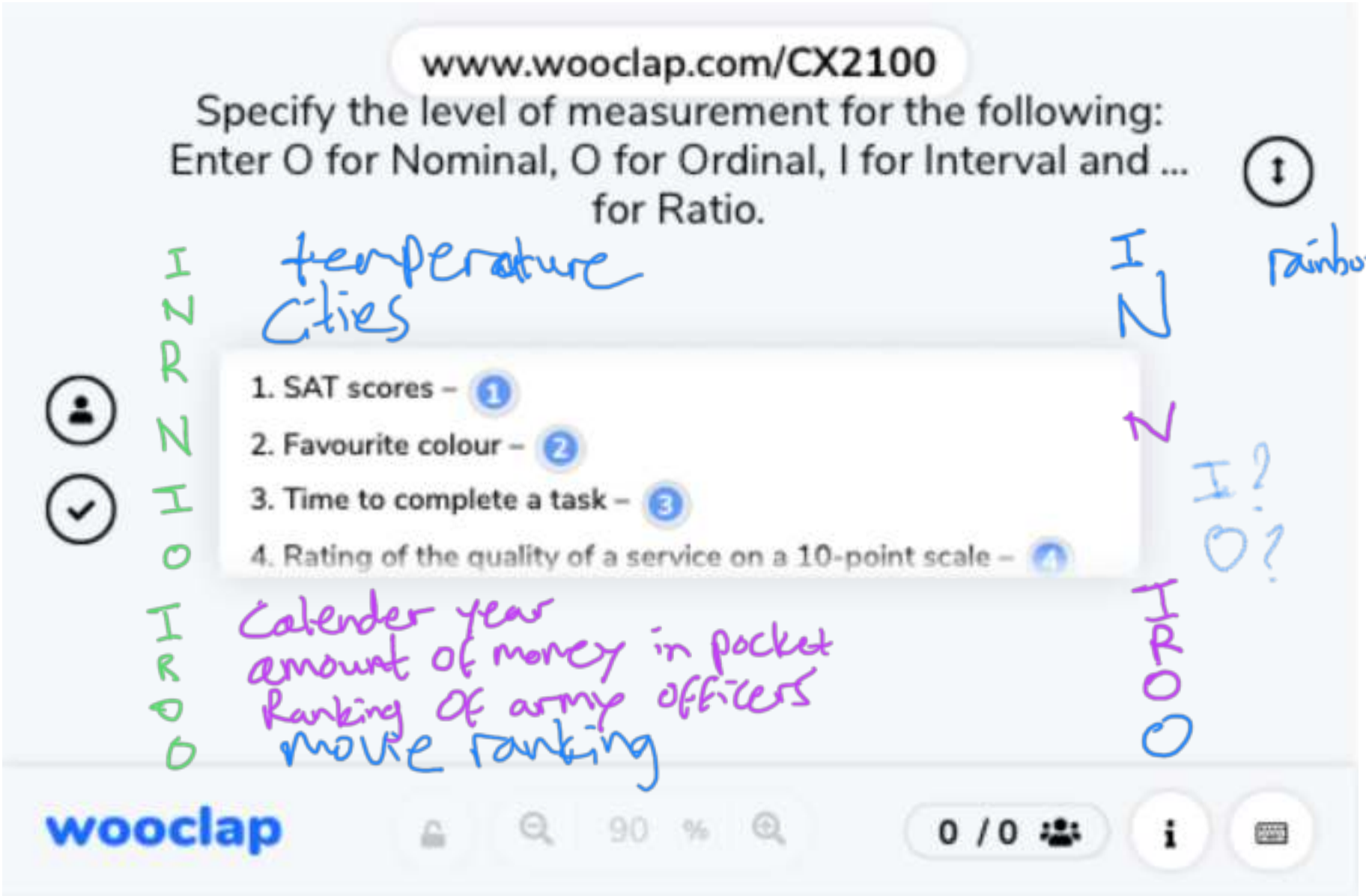
Nominal – names or labels with no specific order

Ordinal – variables in a specific order

Interval – numerical scales in which intervals have the same interpretation throughout, but no true zero

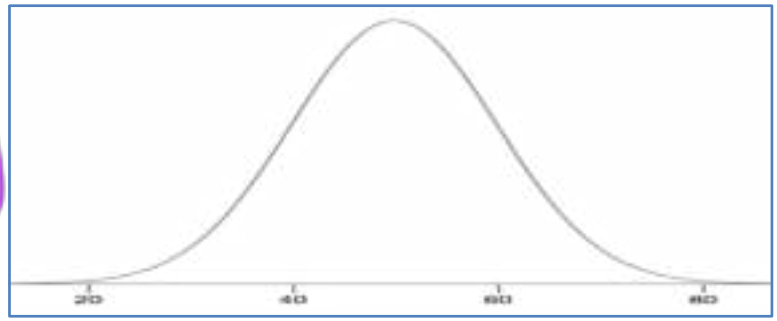
Ratio – include all the characteristics of interval scale, plus it has zero position indicating the absence of the quantity being measured

Specify the level of measurement used for the following variables:



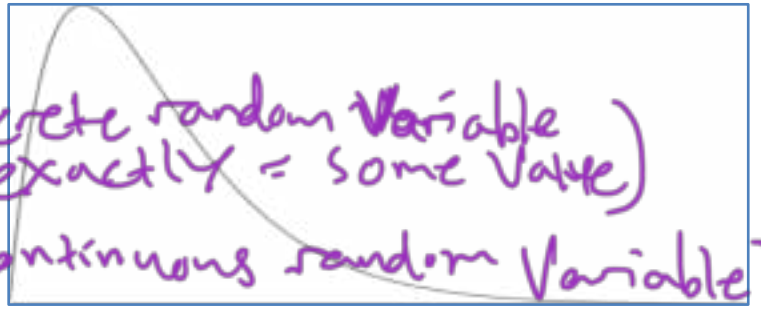
Distributions

Frequency distribution (display in bar chart)
- Discrete variables
- Continuous variables (display in histogram)



Symmetric

Probability distribution
- Probability mass function (discrete random variable is exactly = some value)
- Probability density function (continuous random variable)

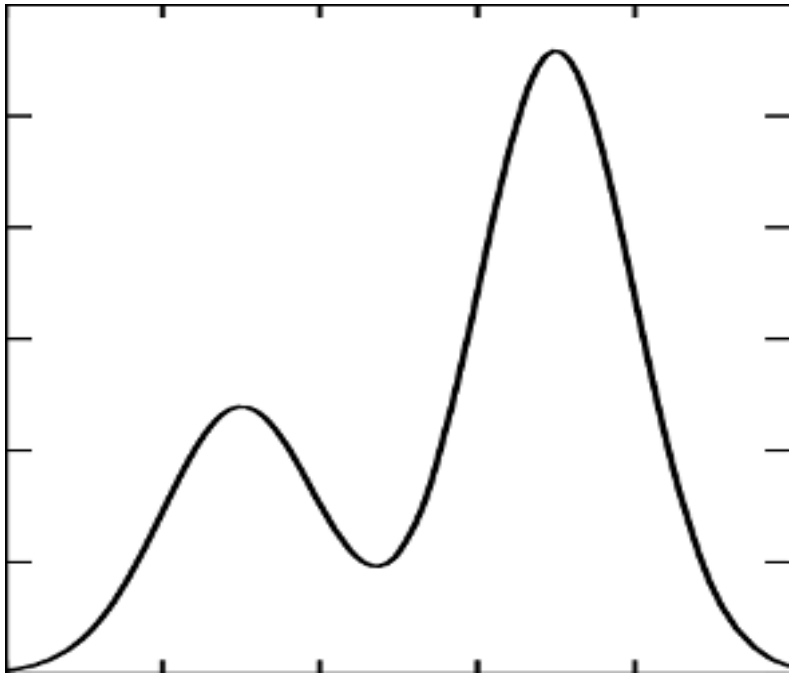


Positive skew (to the right)

Shapes of distributions
- symmetric
- skewed to the right
- skewed to the left

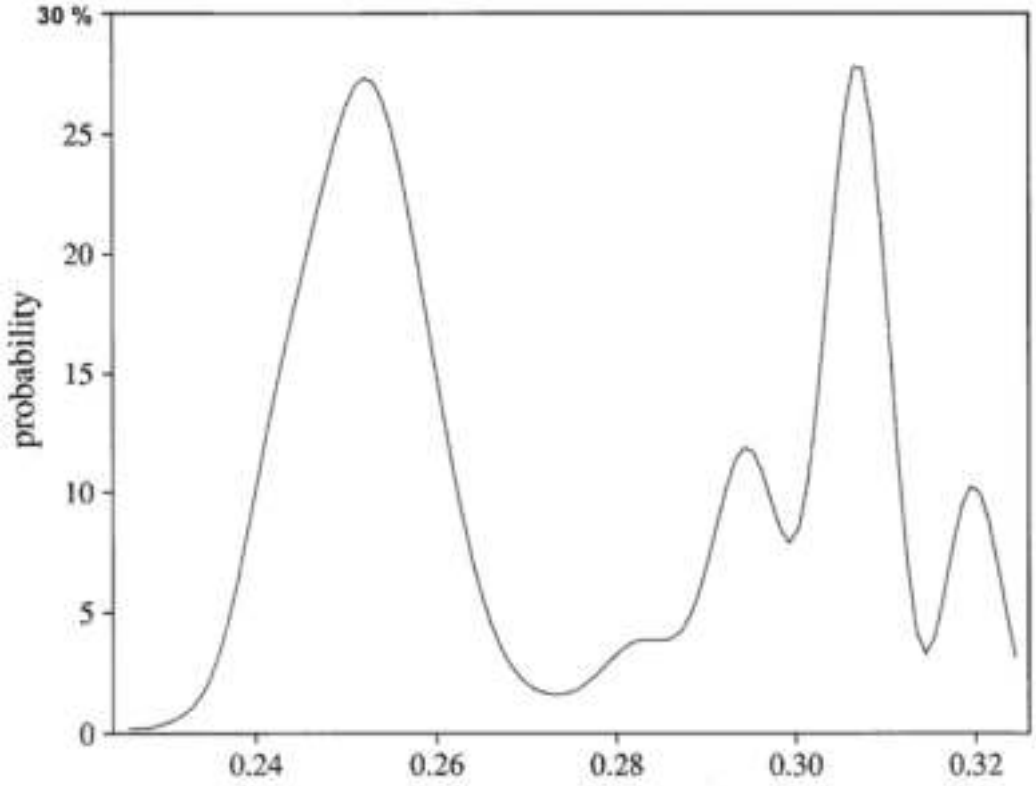
Negative skew (to the left)

- Distributions – other shapes:

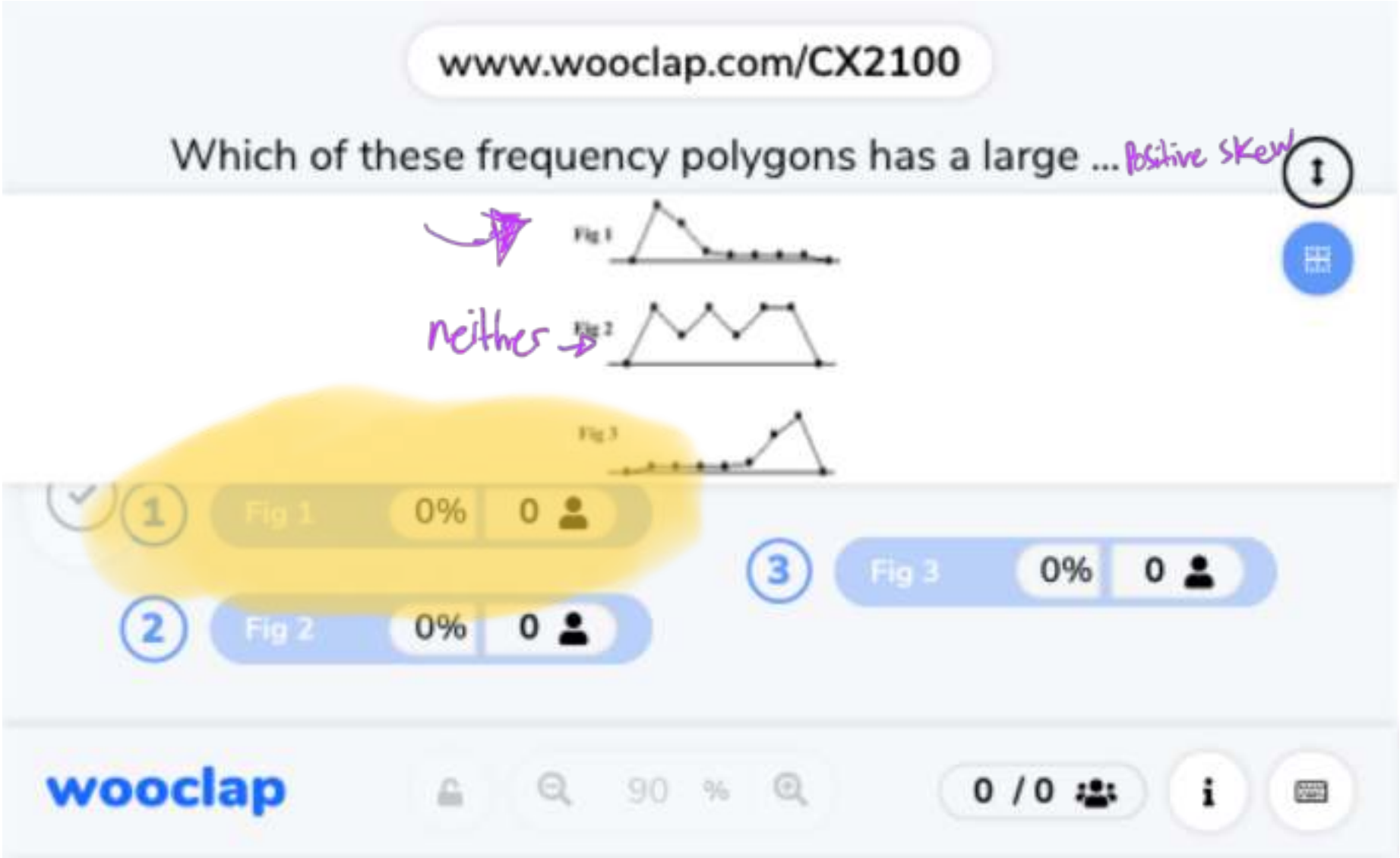


Bimodal distribution

Multi-modal distribution



Which of the following frequency polygons has a large positive skew?



- Linear Transformations

Transform data from one measurement scale to another.

Examples:
Convert length measured in X feet to measurement in Y metre, i.e.

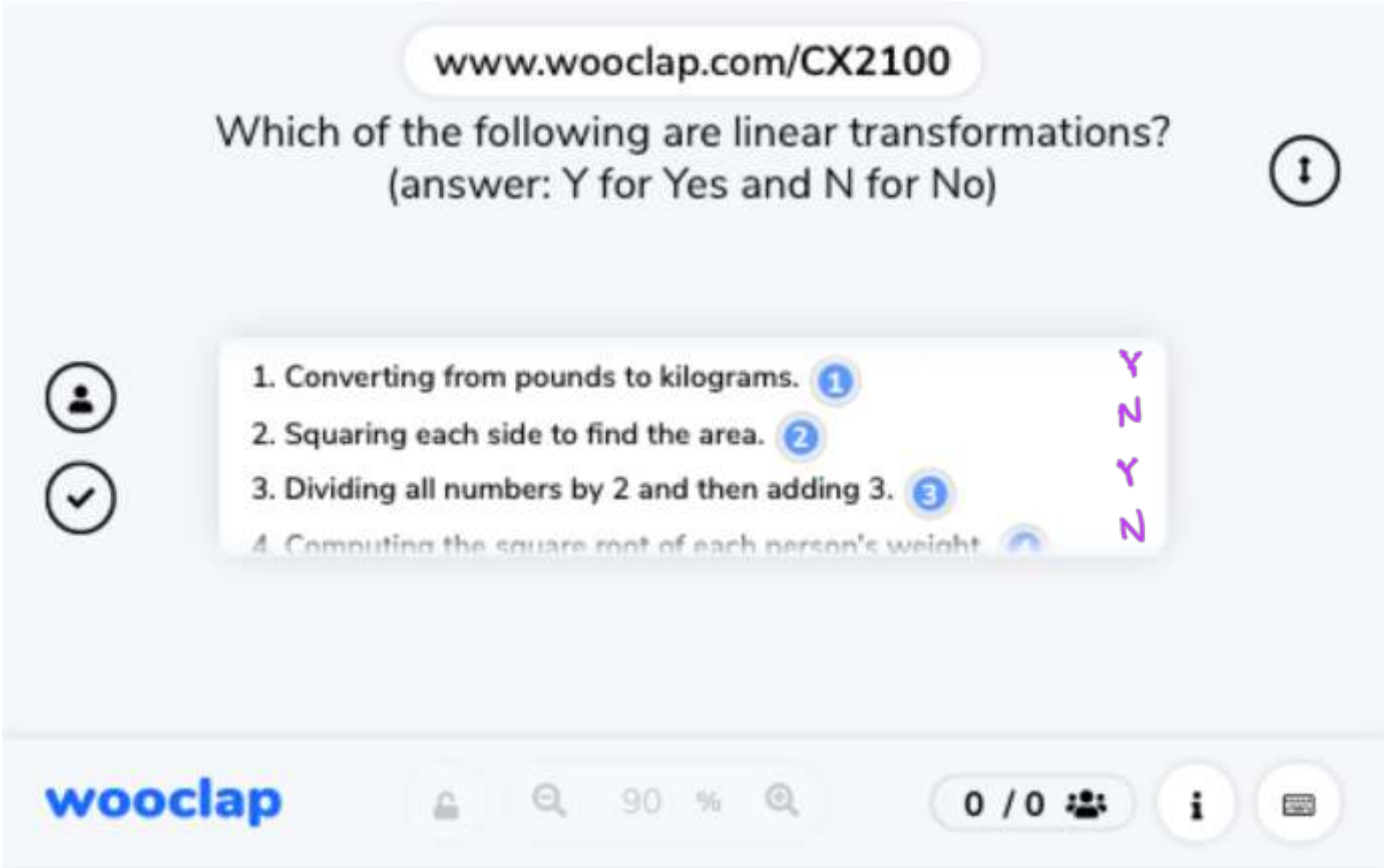
$$Y = 0.3048\,X$$

Convert temperature in Fahrenheit to Centigrade:

$$C = 0.5556\,F\; - 17.778$$

- Linear Transformations

Which of the following are linear transformations?



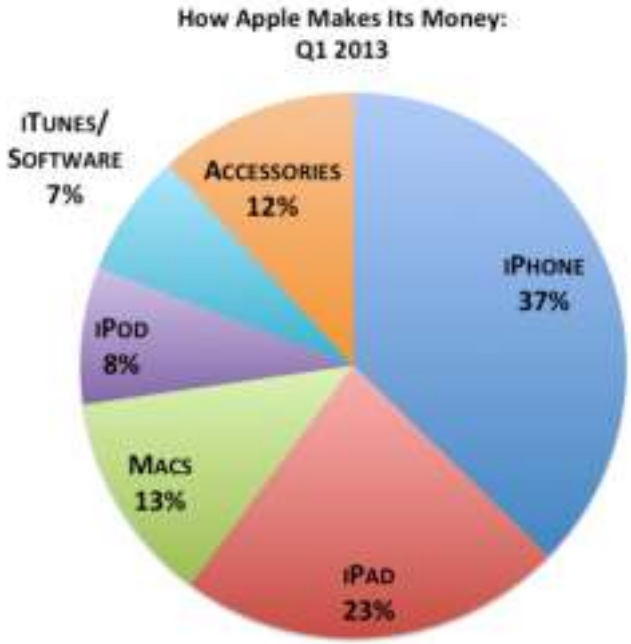
Ch 2. Presenting Data

- Frequency tables and Charts
- Bar Charts
- Stem and Leaf Displays
- Histograms
- Box Plots
- etc

- Presenting qualitative or discrete data:

Previous Ownership	Frequency	Relative Frequency
None	85	0.17
Windows	60	0.12
Macintosh	355	0.71
Total	500	1.00

Frequency table



Bar chart

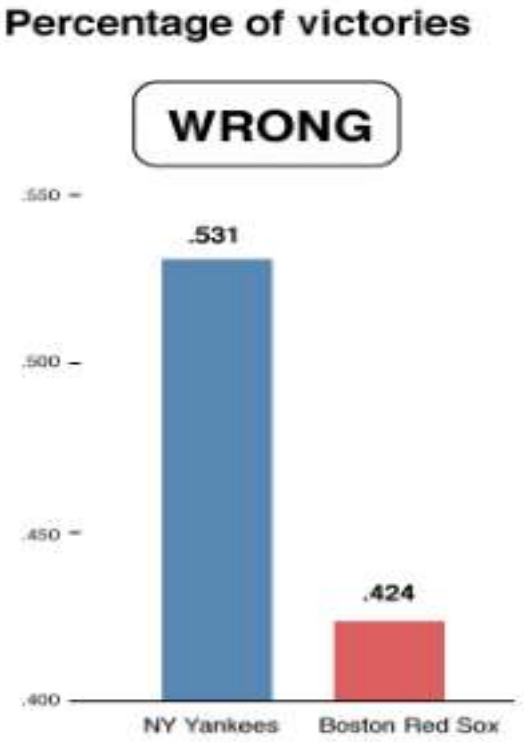
Pie chart

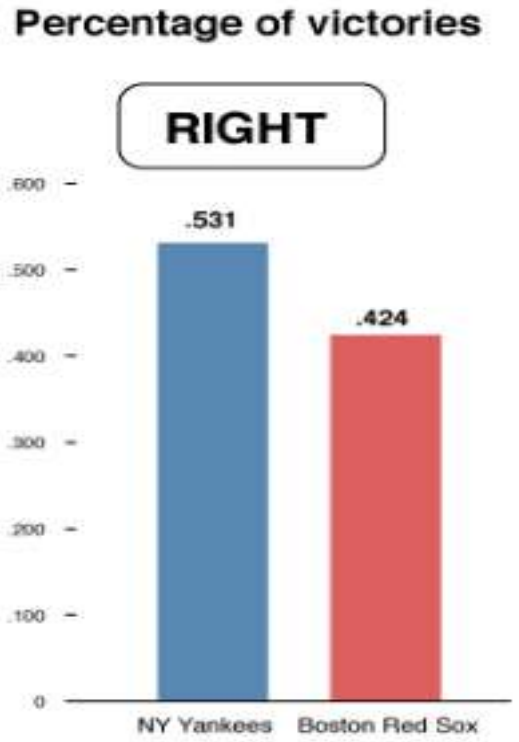
Examples of misleading presentations:

The battery on the right has 70% more capacity than the one on the left.

Misleading display?

yes



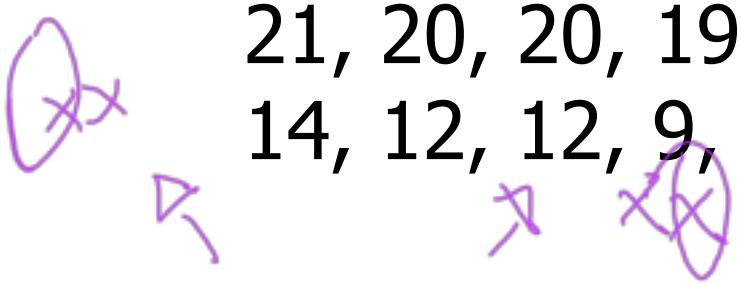


- Stem-and-leaf Presentation

- Useful when data are not too numerous

Eg: No. of touchdown passes by each of the 31 football teams.

37, 33, 33, 32, 29, 28, 28, 23, 22, 22, 22, 21, 21, 21, 20, 20, 19, 19, 18, 18, 18, 18, 16, 15, 14, 14, 14, 12, 12, 9, 6



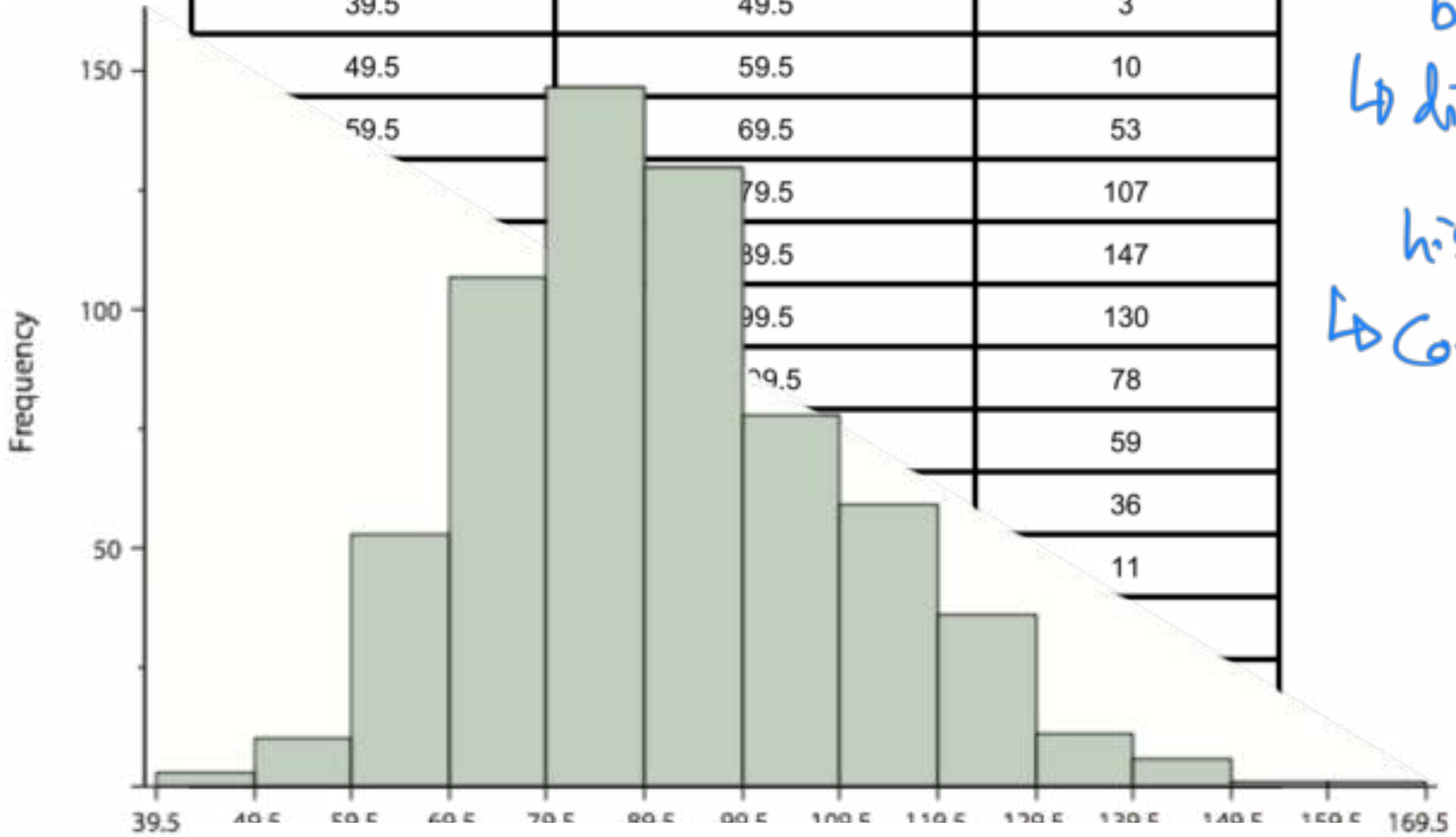
3 | 2337
2 | 001112223889
1 | 2244456888899
0 | 69

Stem Leaf

- Presenting continuous data:

  Data range is divided in class intervals or bins

Interval's Lower Limit	Interval's Upper Limit	Class Frequency
39.5	49.5	3
49.5	59.5	10
59.5	69.5	53
[illegible]79.5	107
[illegible]89.5	147
[illegible]99.5	130
[illegible]09.5	78
		59
		36
		11



Histogram of the test scores

bar chart
↳ discrete data

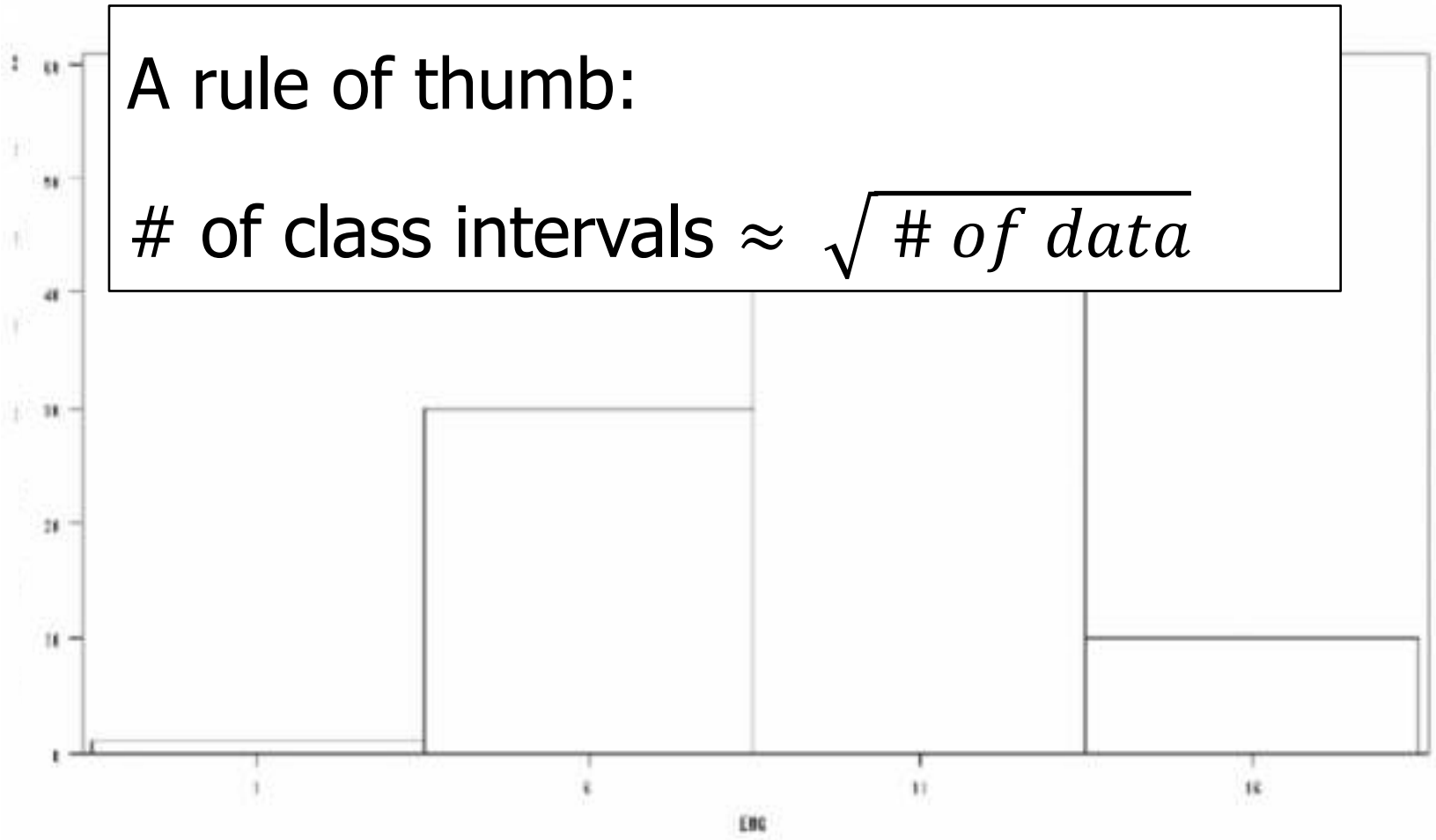
histogram
↳ continuous data

Presenting continuous data - Histogram:

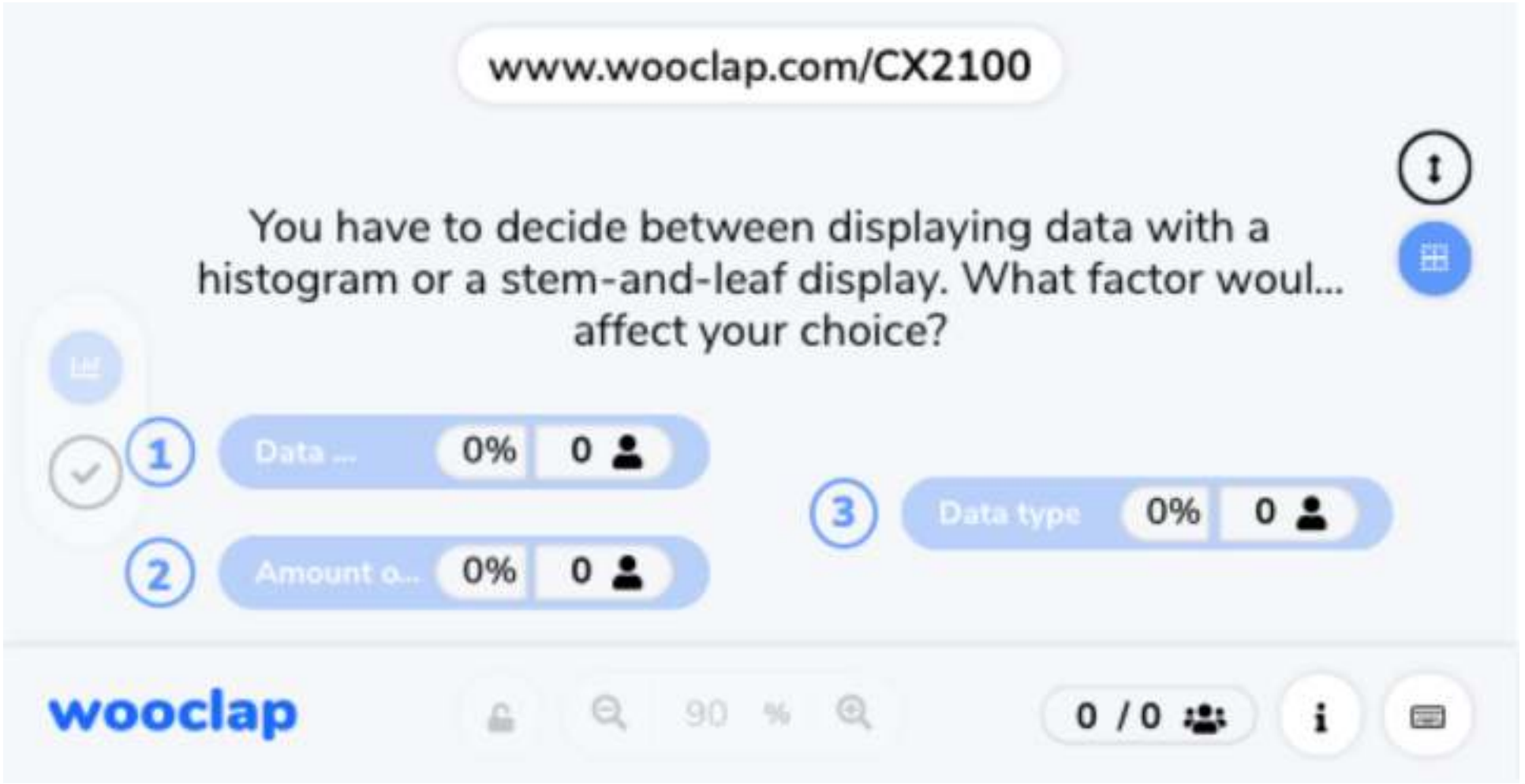
Examples: Class interval size affects the visual presentation

A rule of thumb:

$$\#\ \text{of class intervals} \approx \sqrt{\#\ of\ data}$$



Histogram when class interval is 5



For a given set of data, the choice is to select between a histogram or a stem-and-leaf display.

With more data, a histogram can be very useful since it shows the overall shape of the distribution.

Stem-and-leaf display is better for smaller sets of data.

Question:

Suppose you are constructing a histogram for describing the distribution of salaries for individuals who are 40 years or older, but are not yet retired. (a) What is on the Y-axis? (b) What is on the X-axis? (c) What would be the probable shape of the salary distribution? Explain why.

a) The Y-axis would of the frequency of individuals.

b) Salary would be on the X-axis because this is the variable whose distribution is of interest.

c) The distribution is expected to be positively skewed with a few wealthy CEOs earning well above the average salary.

- Box Plot:

Draw the Box Plot for the following data set (31 values).

14	15	16	16	17	17	17	17	17	18	
18	18	18	18	18	19	19	19	20	20	
20	20	20	20	21	21	22	23	24	24	29

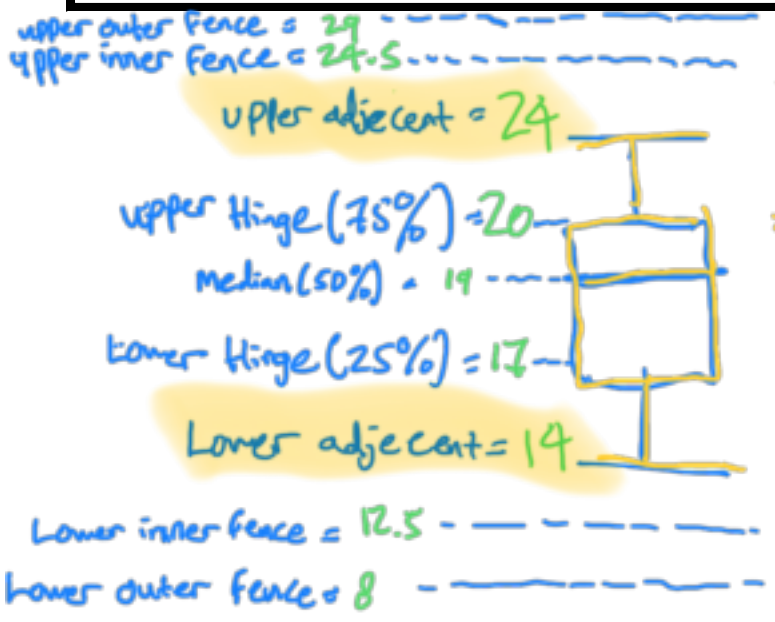
Upper Hinge = 75th Percentile (3rd quartile)	20
Lower Hinge = 25th Percentile (1st quartile)	17
H-Spread = Upper Hinge - Lower Hinge (IQR)	3
Step = 1.5 x H-Spread	4.5
Upper Inner Fence = Upper Hinge + 1 Step	24.5
Lower Inner Fence = Lower Hinge - 1 Step	12.5

(inter quartile range)

Step (1) (2) (3) (4)

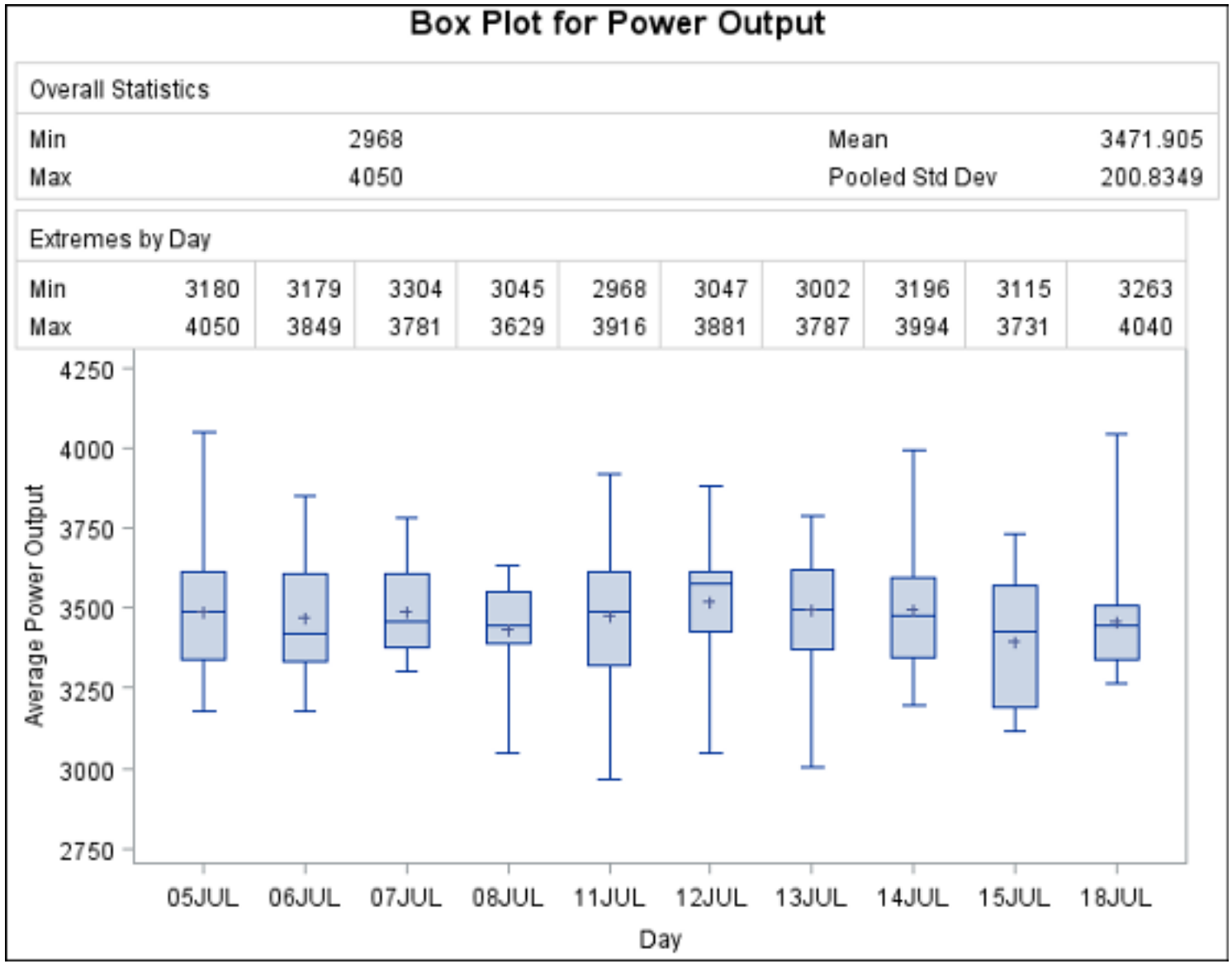
IQR = Upp Hinge − lower Hinge 3

Upper Outer Fence = Upper Hinge + 2 Steps	29
Lower Outer Fence = Lower Hinge - 2 Steps	8
Upper Adjacent = Largest <= Upper Inner Fence	24
Lower Adjacent = Smallest value >= Lower Inner Fence	14
Outside Value — Outlier = Value > upper adjacent or < lower adjacent	29
Far Out Value	



Find 25 - 75 % + 50%
Find IQR / H-Spread
Find step
Find upper/lower inner
Find = adjecent
find upper/lower outer

Practical example of Box Plot:
Daily Power Output of an Electric Generator



Another practical example of Box Plot: Daily share prices

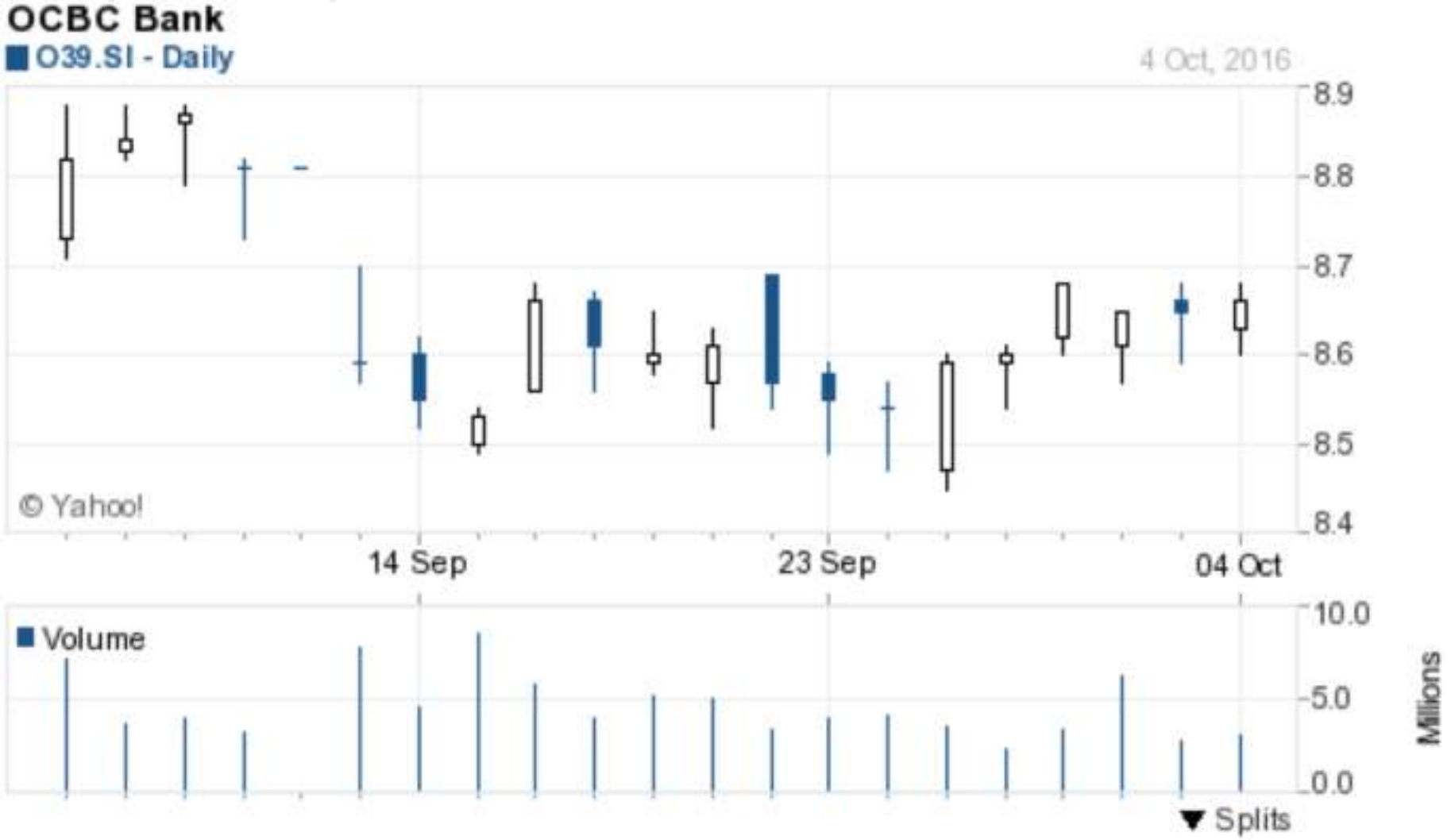
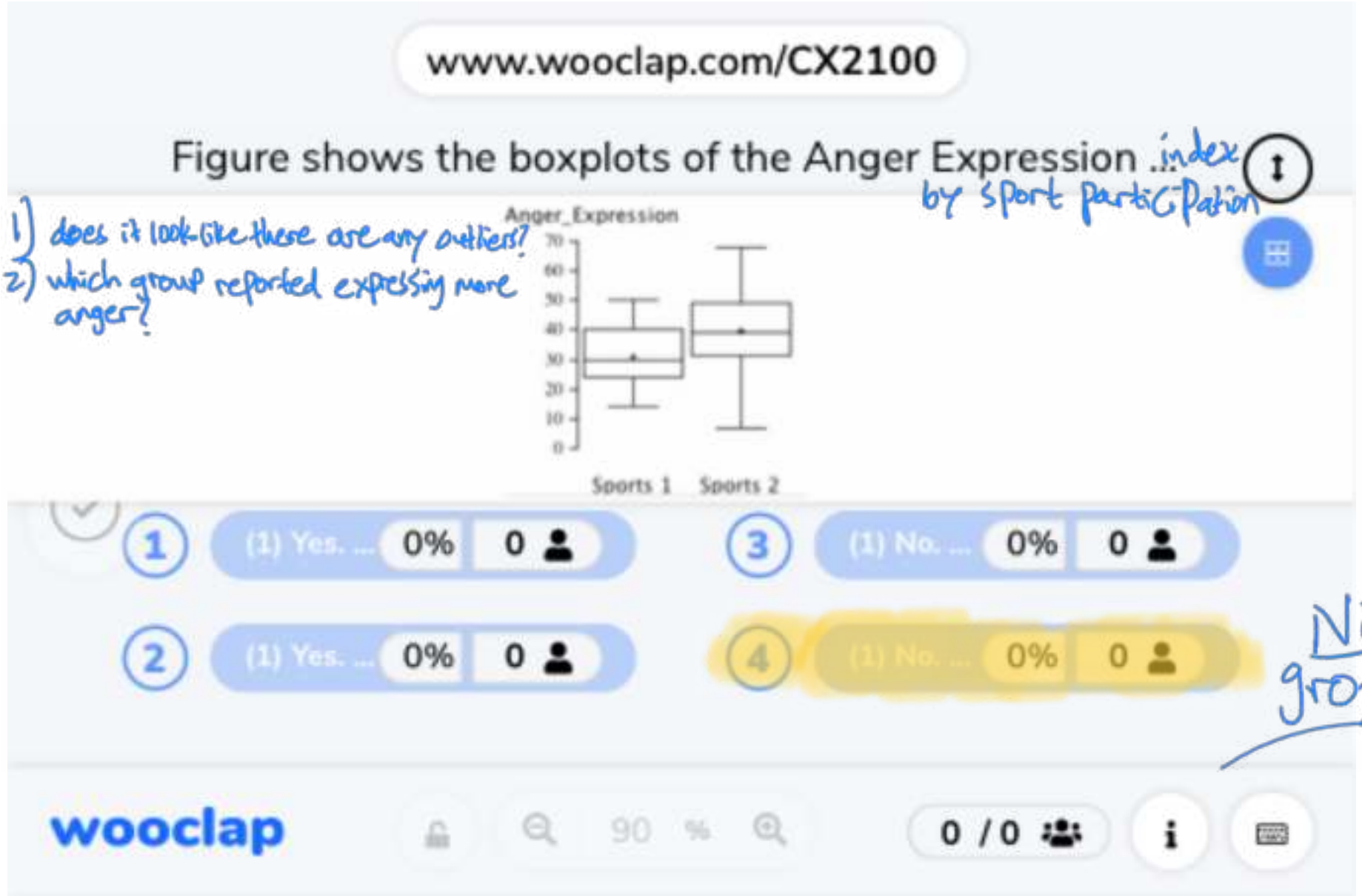


Figure shows the boxplots of the Anger Expression Index by sports participation. Does it look like there are any outliers? Which group reported expressing more anger?



Ch 3. Summarizing Distributions

- Central Tendency: mean, median & mode
- Other Measures of Central Tendency
- Comparing Central Tendency
- Measures of Variability: Range, IQR, Variance
- Linear Transformation of variable
- Variance Sum Law I

- Central Tendency

Summarizes a distribution by its central location

Different ways to define central tendency:

- Mean $\mu = \frac{\sum X}{N}$ → Sum of all value
- Median = 50th Percentile
- Mode = the value with the highest frequency
- Trimean = $\frac{25^{\text{th}}\,\text{Percentile} + 2*\text{Median} + 75^{\text{th}}\,\text{Percentile}}{4}$
- Geometric mean = $(\prod X)^{\frac{1}{N}}$, where $\prod$ means to multiply

  Eg: $\prod_{i=1}^{5} X_i = X_1 \times X_2 \times X_3 \times X_4 \times X_5$

  $\prod_{i=1}^{5} X_i = X_1 * X_2 * X_3 * X_4 * X_5$
- Trimmed mean = mean for data with some higher and lower values removed

- Central Tendency:

Eg: Given the following data set, compute the mean, the median, the mode, the trimean, the geometric mean and the mean trimmed 20%.

1	3	4	4	4	5	5	7	8	9	31

mean = 81/11 = 7.36

median = 5

mode = 4

trimean = (4 + (2×5) + 8)/4 = 5.5

25th = 25/100 × 12 = 3rd → 4

75th = 75/100 × 12 = 9th → 8

Geometric mean

$(\pi x_i)^{1/n}$

$(74\,995\,200)^{1/11} = 5.2$

mean trimmed by 20%

$\frac{3+4+\ldots+9}{7} = 5.7$

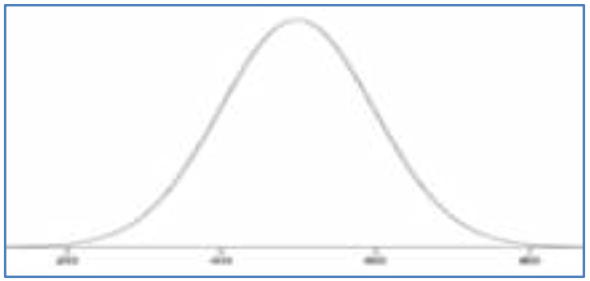
10% of 11 = 1.1 ≈ 1 → remove 1 data each side

- Central Tendency – comparing various measures

For symmetric distributions:

Mean = Median= Trimean = Trimmed mean

= Mode (except bimodal distr)



Symmetric

For skewed distributions:

Differences among the measures.

Example – the mean is typically higher than the median for a positive skewed distribution

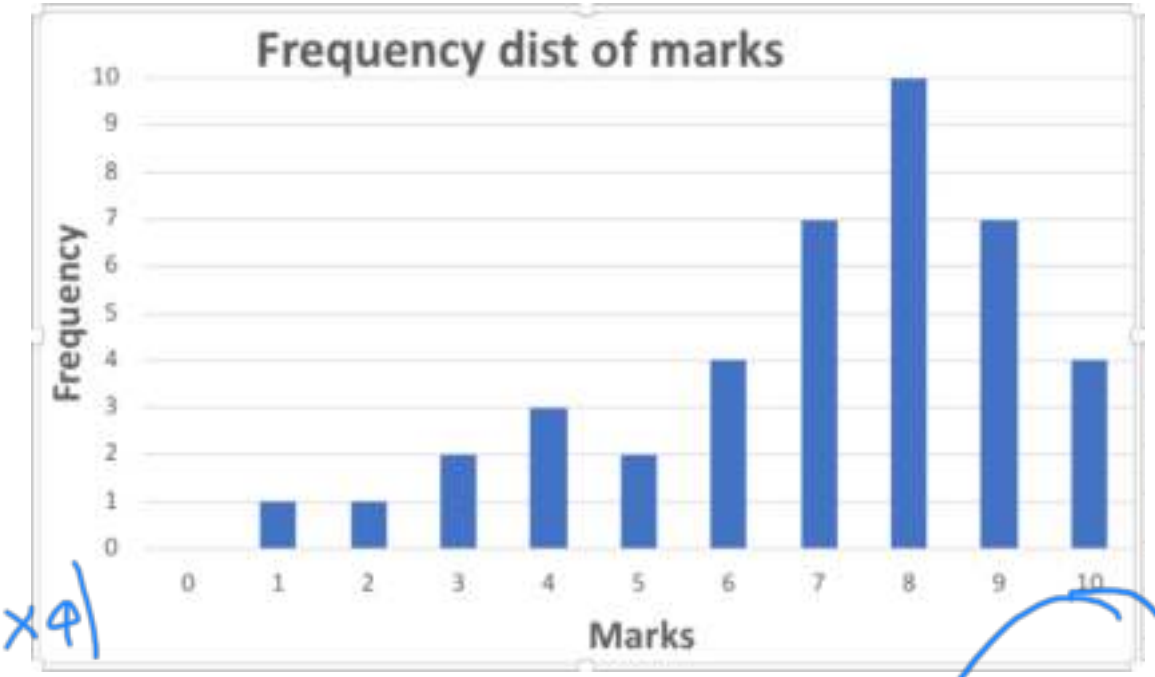
- Central Tendency – comparing various measures

Example:

Negative skewed

Calculate:

1. mode
2. median
3. trimean
4. mean



mode = 8

mean = $\frac{(1\times1)+(2\times1)+(3\times2)+\cdots+(10\times4)}{41} = 7$

median = midpoint = 8

trimean = $\frac{6+2\times8+9}{4} = \frac{31}{4} = 7.75$

25th Percentile 10.5 → 6

75th Percentile 31.5 → 9

- Measures of Variability

An indication of how spread out is the distribution

Frequently used measures of variability:

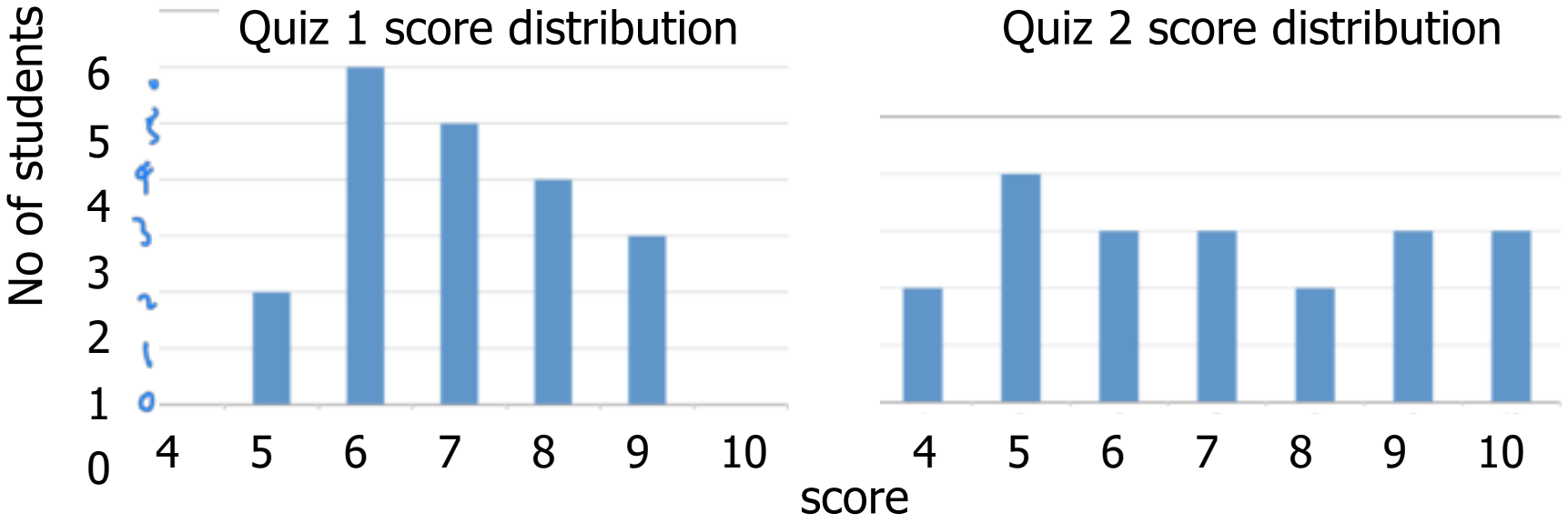
- Range = Highest value – Lowest value
- Interquartile Range IQR = 75th – 25th Percentile
- Variance $\sigma^2 = \frac{\sum(X - \mu)^2}{N}$
- Standard deviation = $\sqrt{\text{Variance}}$

Each data mean

Variance must be +

- # Variability:

Eg: Given the following 2 data set, compute the mean, the range, the IQR and the variance.



Mean μ	7	7
Range	$9-5=4$	$10-4=6$
IQR	$8-6=2$	4
Var σ^2	$\sigma^2 = \frac{\Sigma(x-\mu)^2}{20} = 1.5$	3.9

$2(5-7)^2 + 6\cdot(6-7)^2 + 5(7-7)^2 + 4(8-7)^2 + 3(1-7)^2/20$

$\frac{8+6+0+4+12}{20} = 1.5$

$E[x^2] = \frac{\Sigma x^2}{N} = (\sigma^2 + \mu^2)N$ | $n\mu = E[\Sigma x] = \Sigma E[x]$ $n\sigma^2 = Var[\Sigma x] = \Sigma Var[x]$

Compute mean and variance from the **population** of size *N* :

Population Mean $\mu = E[X] = \frac{\sum X}{N}$

Population Variance $\sigma^2 = E[(X-\mu)^2] = \frac{\sum (X-\mu)^2}{N}$ or $\frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N}$

$\sigma^2 = E[X^2] - \mu^2$

Estimate mean and variance from a **sample** of size *n* :

Sample Mean $\bar{x} = \frac{\sum X}{n}$

Sample Variance $s^2 = \frac{\sum (X-\bar{x})^2}{n-1}$ or $\frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n-1}$

Why n-1 ?

Suppose the denominator of s^2 is n, instead of $(n-1)$:

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

For unbiased estimate, we expect the mean of s^2 to be equal to σ^2 :

$$E[s^2] = E\left[\frac{1}{n}\left(\sum X^2 - \frac{(\sum X)^2}{n}\right)\right]$$

$$= \frac{1}{n}\left(\sum E[X^2] - \frac{E[(\sum X)^2]}{n}\right)$$

$\sigma^2 + \mu^2$ from the previous slide

$$= \frac{1}{n}\left(n\,\sigma^2 + n\,\mu^2 - \frac{E[(\sum X)^2]}{n}\right)$$

$$E[s^2] = \frac{1}{n}\left(n\,\sigma^2 + n\,\mu^2 - \frac{E[(\sum X)^2]}{n}\right)$$

Let $Y = \sum X$

$E[Y^2] = \sigma_Y^2 + \mu_Y^2$ → $E[\sum X]^2$

→ $Var[\sum X]$

$$= \frac{1}{n}\left(n\,\sigma^2 + n\,\mu^2 - \frac{Var[\sum X] + (E[\sum X])^2}{n}\right)$$

$$= \frac{1}{n}\left(n\,\sigma^2 + n\,\mu^2 - \frac{\sum Var[X] + (\sum E[X])^2}{n}\right)$$

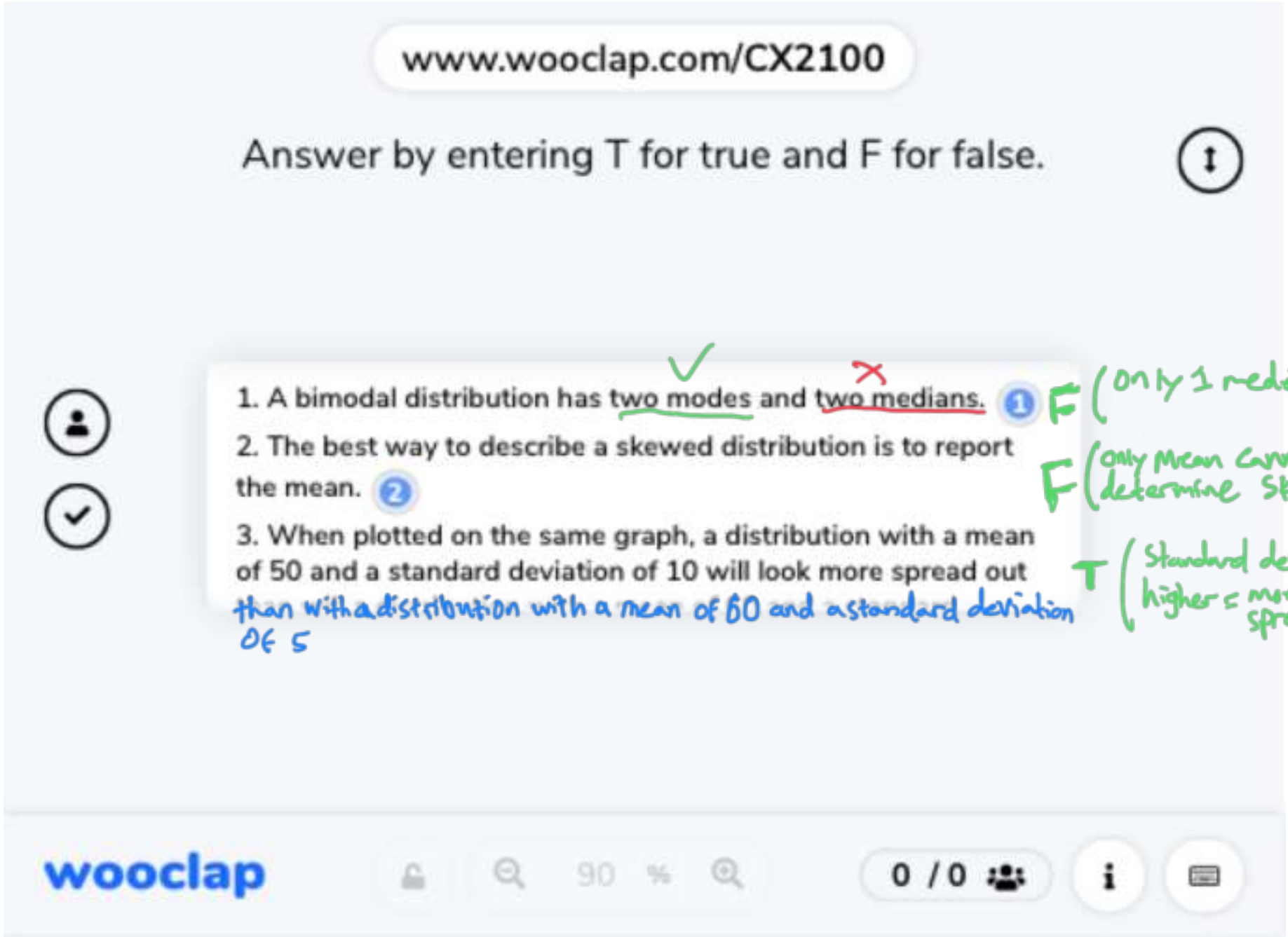
$$= \frac{1}{n}\left(n\,\sigma^2 + n\,\mu^2 - \frac{n\,\sigma^2 + (n\,\mu)^2}{n}\right)$$

$$= \frac{1}{n}(n\,\sigma^2 + n\,\mu^2 - \sigma^2 - n\,\mu^2) \quad = \frac{1}{n}(n-1)\,\sigma^2$$

If the denominator is $(n-1)$,

then mean of s^2 is equal to σ^2, i.e. $E[s^2] = \sigma^2$

True or False questions:



- Linear transformation of variable

  - Transform data from one measurement scale to another

Eg. Consider a taxi trip from point A to B. The taxi service initial charge is $3 and additional $0.50 per km for the trip. Let y be the cost of the taxi ride and x be the distance travelled. We have:

$$y = 0.5x + 3$$

The mean of y = 0.5 (mean of x) + 3

The variance of $y = 0.5^2$ (variance of x)

- Variance Sum Law I

- Linear combination of 2 independent variables

Eg. Consider a couple A and B. A works x hrs/day and earns \$5/hr while B works y hrs/day and earns \$10/hr. Total daily income t is:

$$t = 5x + 10y$$

Daily income difference d:

$$d = 5x - 10y$$

Mean = 5 (mean of x) − 10 (mean of y)

Variance = $5^2\ \sigma_x^2 + 10^2\ \sigma_y^2$

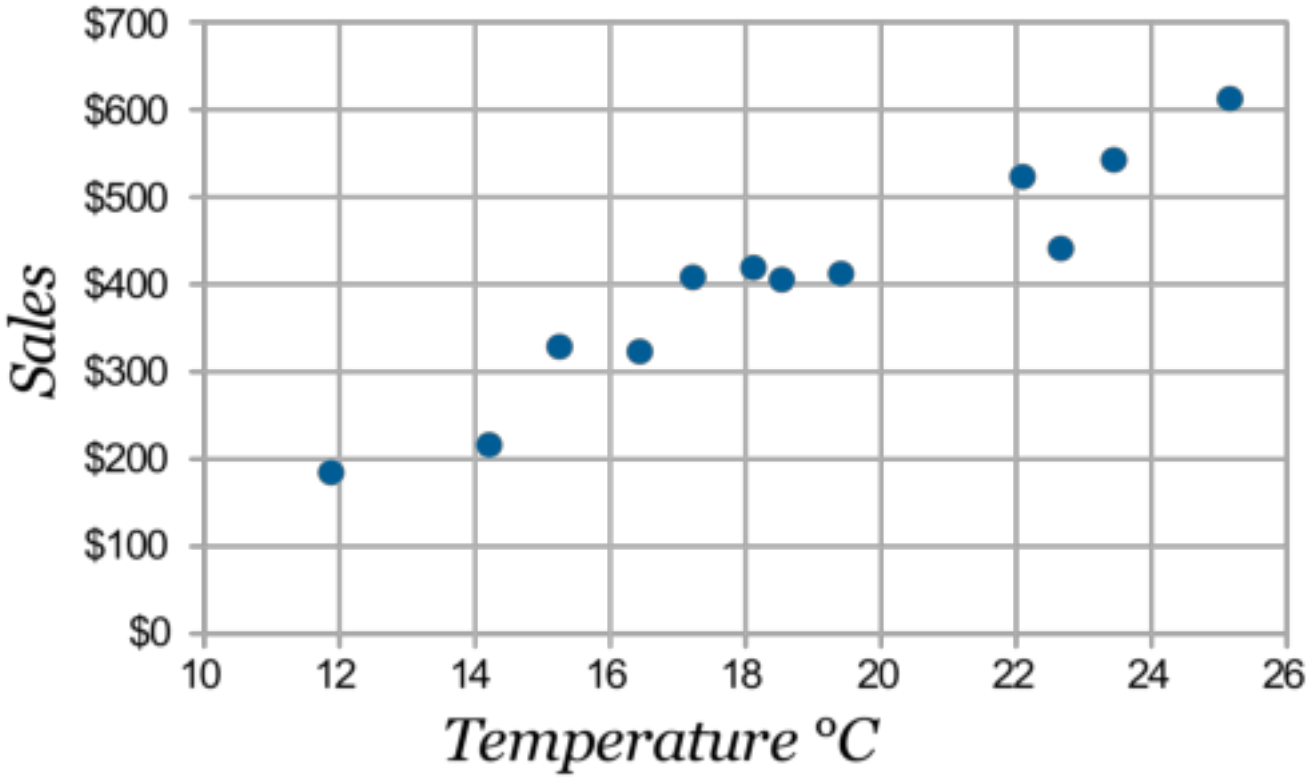
Ch 4. Bivariate Data

- Introduction to Bivariate Data
- Pearson Correlation and Covariance
- Properties of ~~Person~~ Pearson Correlation
- Variance Sum Law II

- Bivariate Data

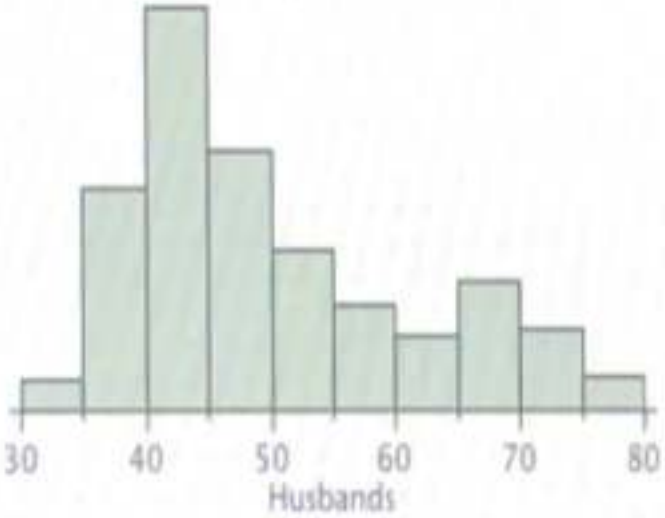
  - A dataset with a pair of variables which may be correlated to one another.

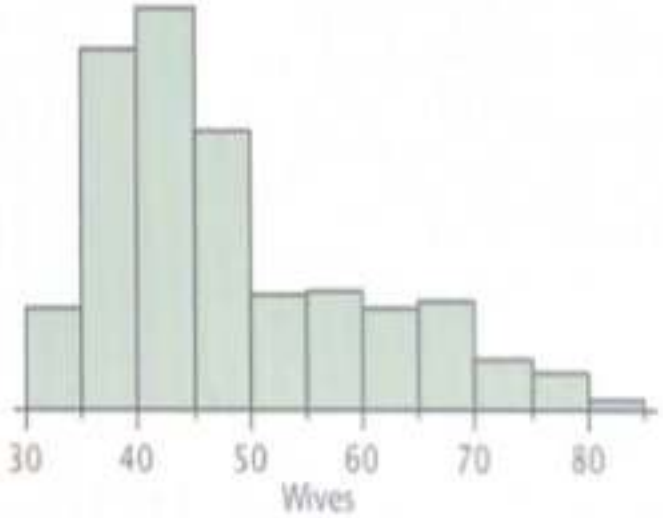
Eg: two variables – ice cream sales and temperature



- Bivariate Data

Eg: Figures below show the frequency distribution of the age of husband and wife for 280 couples.



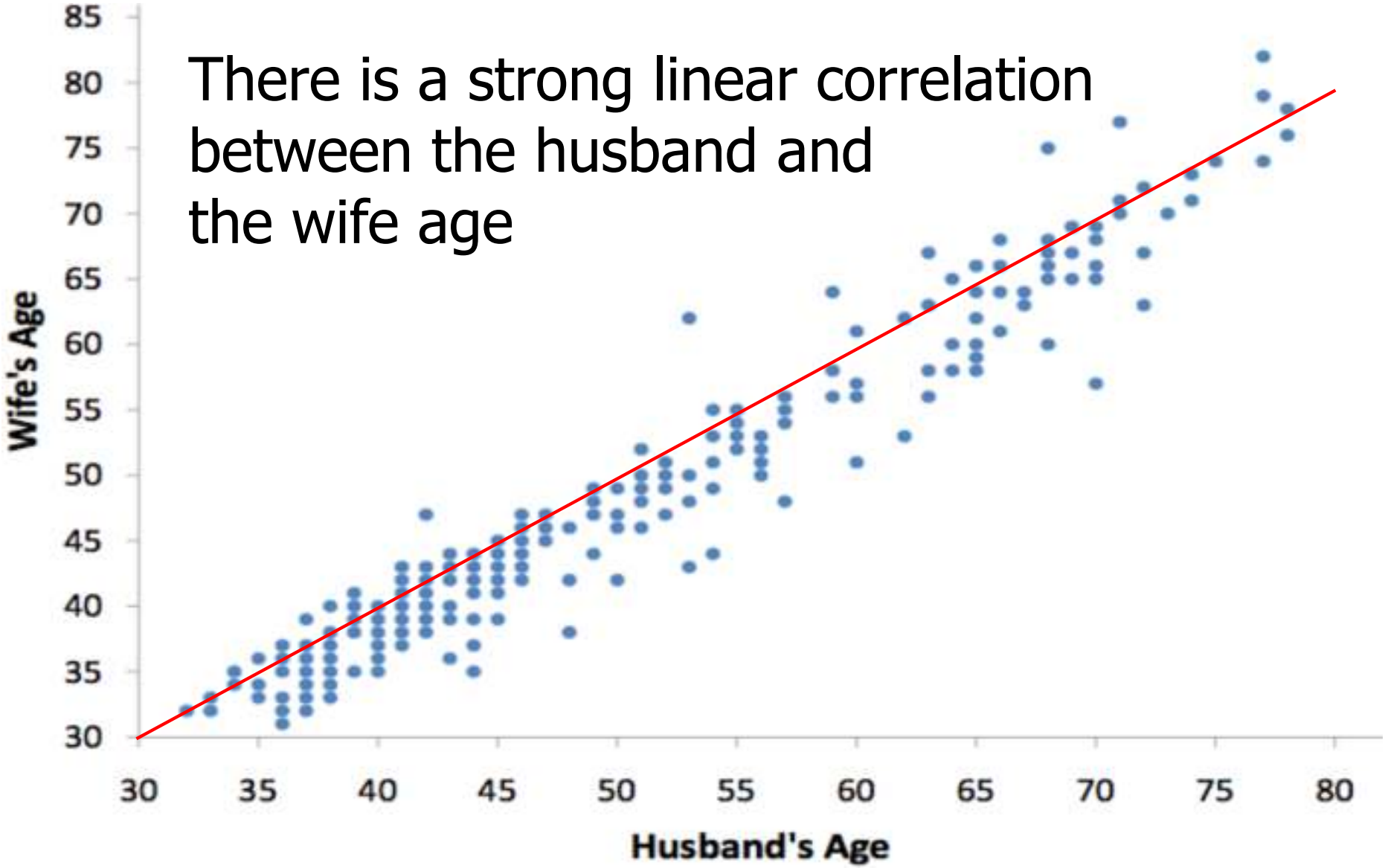


Indication on the central tendency, variability and the shape of distribution of husband and wife age.

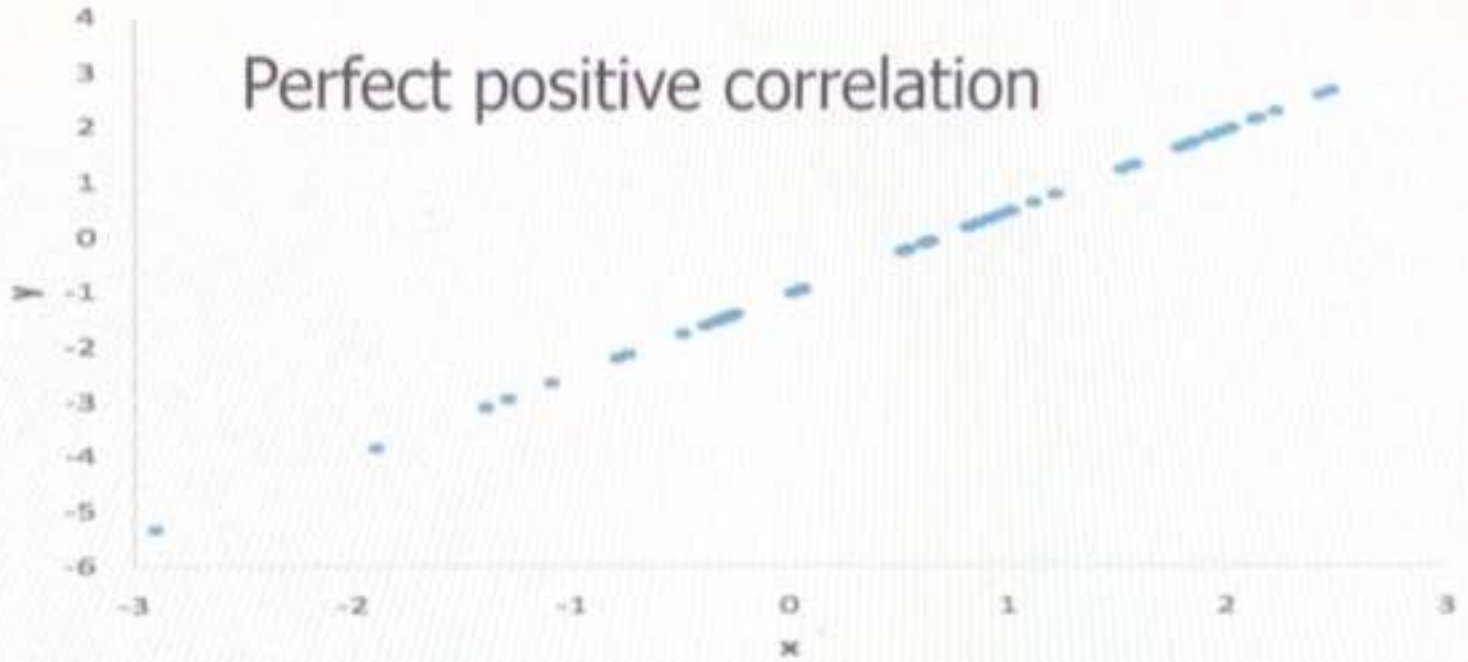
No indication on the correlation of the 2 variables.

Use this one

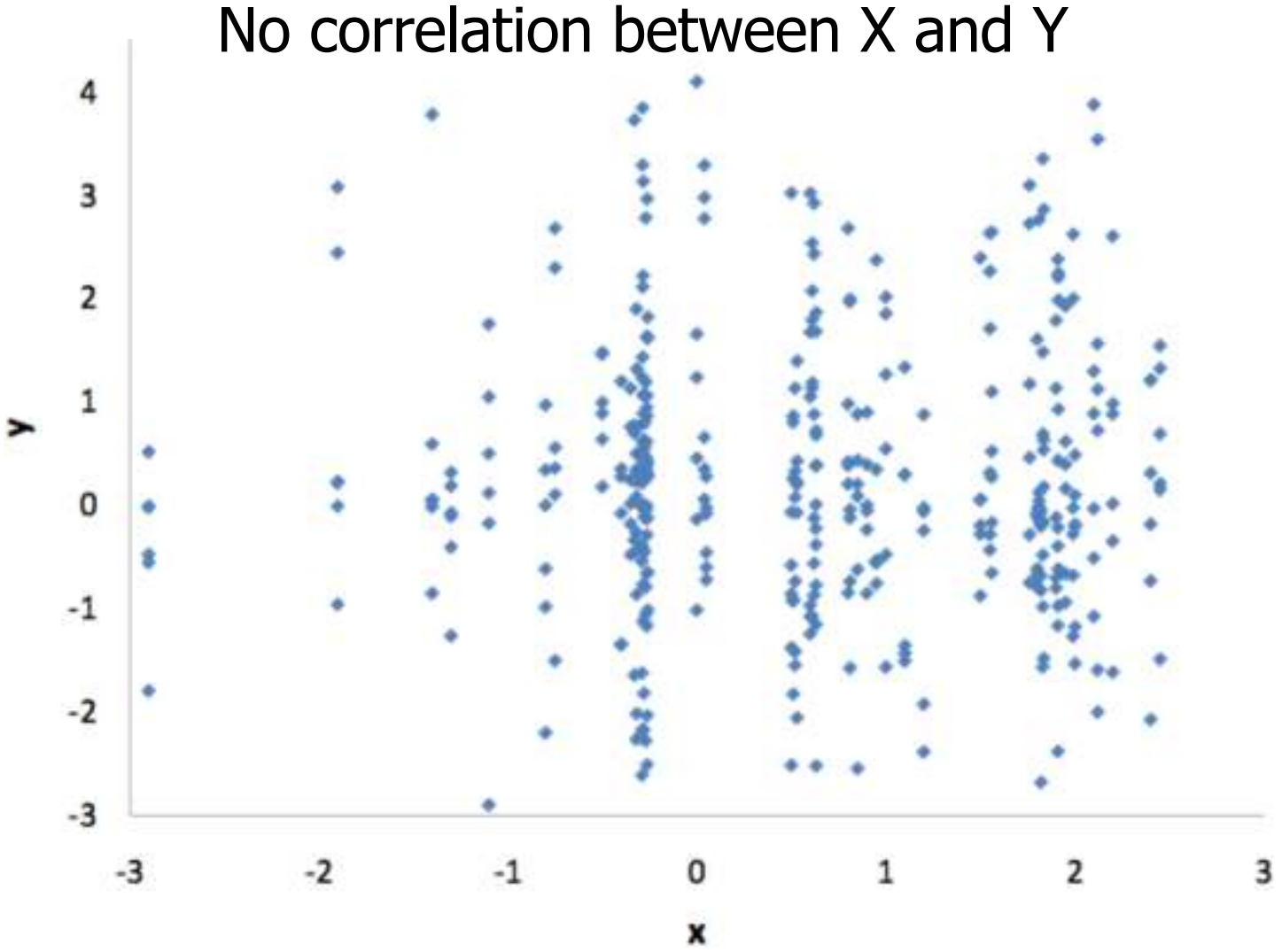
- Bivariate Data



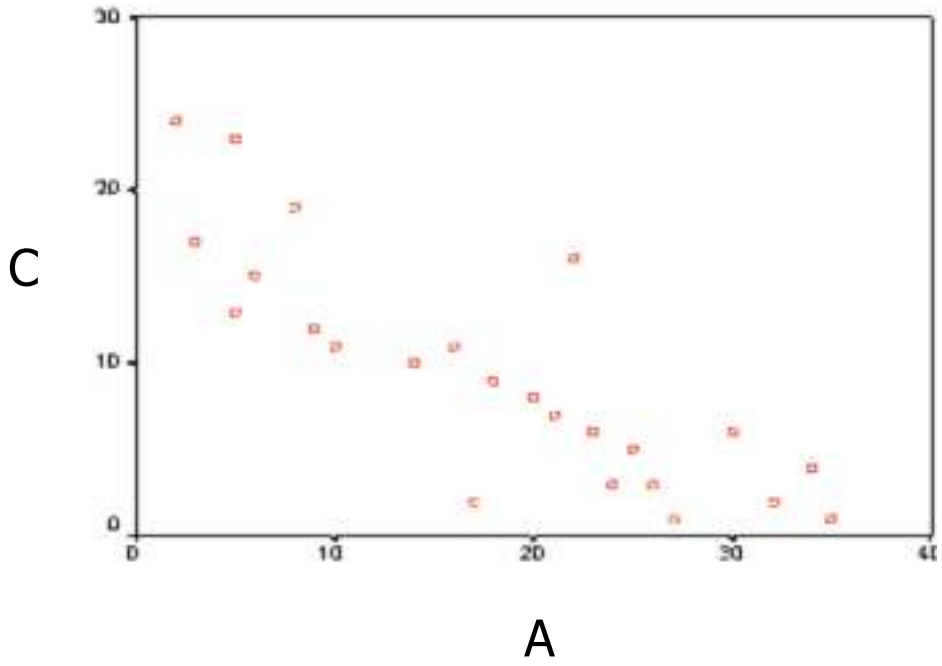
Eg: Correlation of variables X and Y



Eg: Correlation of variables X and Y



Question: Describe the relationship between variables A and C. Think of things these variables could represent in real life.



Negative relationship between A and C.
There is a negative relationship between price and quantity of the products that we buy.

Pearson Correlation ρ

! Population

- An indicator on the strength of the linear relationship between two variables.

distance of X = $X - \mu_X$

distance of Y = $Y - \mu_Y$

Definition: $\rho = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$

Covariance of X and Y, denoted as cov(XY)

$= E[(X-\mu_X)(Y-\mu_Y)]$

$$= \frac{E[XY]-\mu_X\mu_Y}{\sqrt{E[X^2]-(\mu_X)^2}\sqrt{E[Y^2]-(\mu_Y)^2}}$$

$$= \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\sum X^2 - \frac{(\sum X)^2}{N}}\sqrt{\sum Y^2 - \frac{(\sum Y)^2}{N}}}$$

If $\mu_X = \mu_Y = 0$, then $\rho = \frac{\sum XY}{\sqrt{\sum X^2}\sqrt{\sum Y^2}}$

- Computation of Correlation based on a sample of size n

$$\text{cov}(XY) = \frac{1}{n-1}\sum (X-\bar{X})(Y-\bar{Y})$$

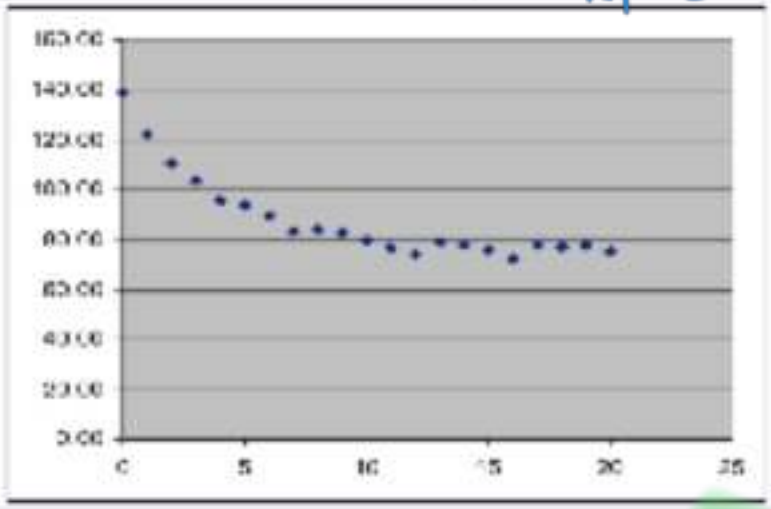
$$\text{Correlation } r = \frac{E[(X-\bar{X})(Y-\bar{Y})]}{s_X s_Y}$$

$$= \frac{\sum XY - \frac{\sum X \sum Y}{n}}{\sqrt{\sum X^2 - \frac{(\sum X)^2}{n}}\sqrt{\sum Y^2 - \frac{(\sum Y)^2}{n}}}$$

$$\text{If } \bar{X} = \bar{Y} = 0, \text{ then } r = \frac{\sum XY}{\sqrt{\sum X^2}\sqrt{\sum Y^2}}$$

www.wooclap.com/CX2100

Given the data shown in the figure, is it appropriate to... use Pearson correlation to describe the relationship between x and y ?



1 Yes 0% 0

2 No 0% 0

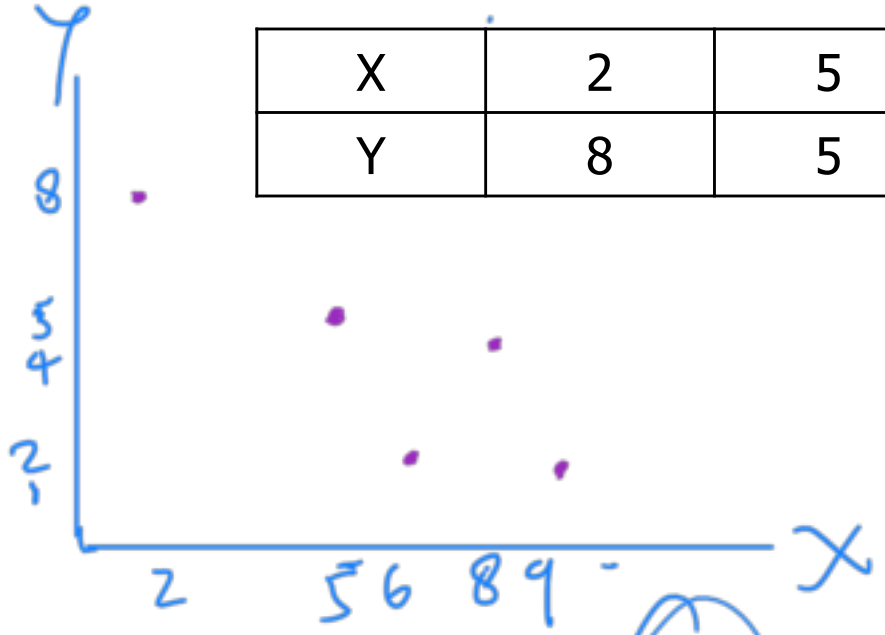
because it is not a linear curve

wooclap 90 % 0 / 0

$\sqrt{6}$ $\sqrt{6}$ -5.2 -0.867

Eg: Given the data set, calculate σ_X, σ_Y, cov(X,Y) and the Pearson Correlation.

X	2	5	6	8	9
Y	8	5	2	4	1



$$\mu_X = \frac{\Sigma X}{N} = \frac{30}{5} = 6$$

$$\mu_Y = \frac{\Sigma Y}{N} = \frac{20}{5} = 4$$

$$\sigma_X^2 = \frac{\Sigma (X-\mu_X)^2}{N} = \frac{16+1+0+4+9}{5} = 30/5 = 6$$

$$\sigma_Y^2 = \frac{\Sigma (Y-\mu_Y)^2}{N} = \frac{16+1+4+0+9}{5} = 6$$

$$\text{cov}(X,Y) = \frac{\Sigma (X-\mu_X)(Y-\mu_Y)}{N} = \frac{-16-1+0+0-9}{5} = -26/5 = -5.2$$

$$\rho = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{-5.2}{\sqrt{6}\cdot\sqrt{6}} = -0.867$$

Properties of Correlation

- Value in the range of [−1, +1]

- Symmetric: correlation of X with Y
  = correlation of Y with X

- Unaffected by linear transformations:

  Correlation of Y with X
  = correlation of Y with A X + B
  where A and B are constants

Eg: If the correlation between weight (in pounds) and height (in feet) is 0.58, find:

(a) the correlation between weight (in pounds) and height (in yards)

(b) the correlation between weight (in kilograms) and height (in meters).

The correlation for both (a) and (b) is still 0.58 because linear transformations do not affect the value of Pearson's correlation, and both of the above instances are linear transformations.

Variance Sum Law II

- Linear combination of 2 independent variables X and Y

  Variance of $X \pm Y$: $\sigma^2_{x\pm y} = \sigma^2_x + \sigma^2_y$

- If the variables X and Y are correlated

  Variance of $X \pm Y$: $\sigma^2_{x\pm y} = \sigma^2_x + \sigma^2_y \pm 2\rho\sigma_x\sigma_y$

- For computation based on a sample

$$s^2_{x\pm y} = s^2_x + s^2_y \pm 2rs_xs_y$$

Eg: Students took 2 parts of a test, each worth 50 points. Part A has a variance of 25, and Part B has a variance of 49. The correlation between the test scores is 0.6.

(a) If the teacher adds the grades of the two parts together to form a final test grade, what would the variance of the final test grades be?

(b) What would the variance of Part A - Part B be?

$\sigma_a^2 = 25$

$\sigma_b^2 = 49$

$\rho = 0.6$

a) $\sigma^2_{x+y} = \sigma_x^2 + \sigma_y^2 + 2\rho\sigma_x\sigma_y$ (adds)

$= 25 + 49 + 2 \cdot 0.6 \cdot 5 \cdot 7$

$= 74 + 42 = 116$

b) $\sigma^2_{x-y} = 74 - 42 = 32$ (Part A - Part B)

Ch 5. Probability

- Basic Concepts – Review of Set Theory and Venn Diagram
- Counting – Ordered, Unordered, Sampling With and Without Replacement.
- Probability Theory
- Base Rate: Bayes' Theorem

- Set Theory - Review

A set is a collection of elements.

Eg: A = {Head, Tail}, Head ∈ A, Tail ∈ A

B = {1, 2, 3, 4, 5, 6}

C = {1, 3, 5}

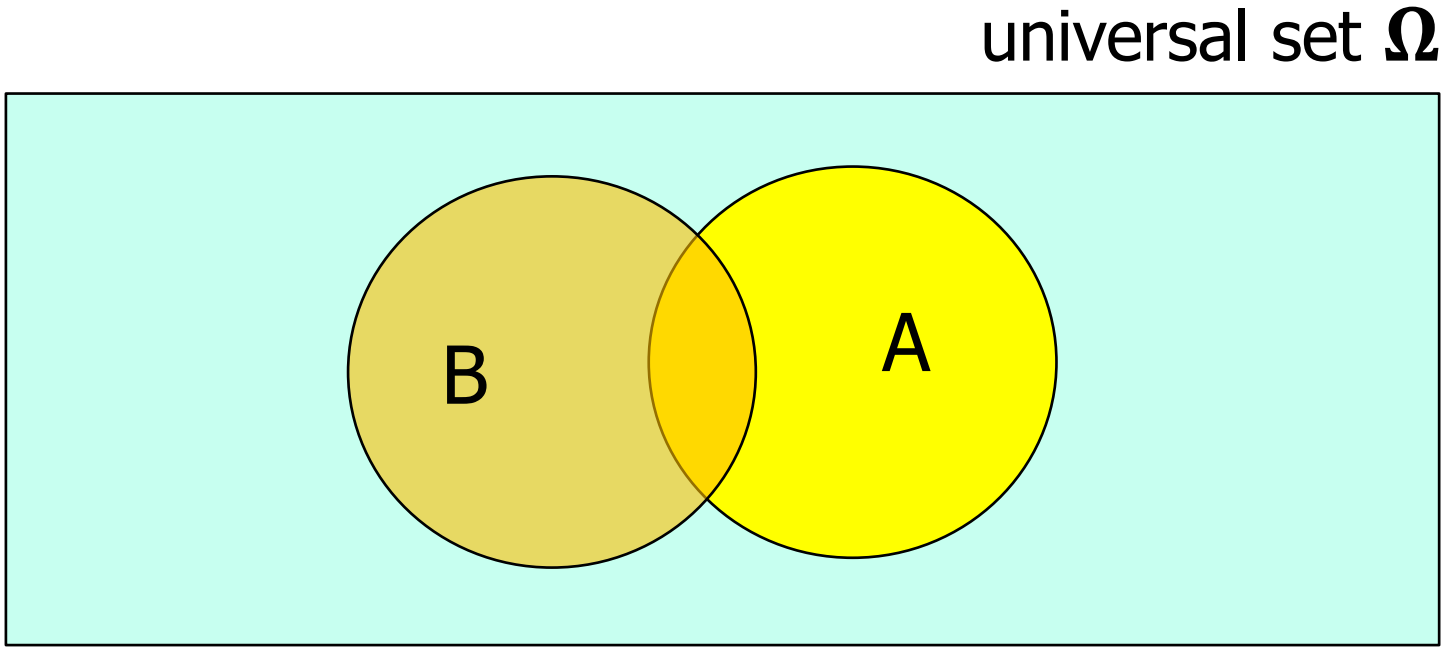
Set C is a subset of Set B since all elements in C are also in B. Notation: C ⊂ B

The set of all elements in a given context is known as an universal set **Ω**. = Set of all Elements

Null set 0 is a set with no element.

A' : Complement of A

- Venn Diagram

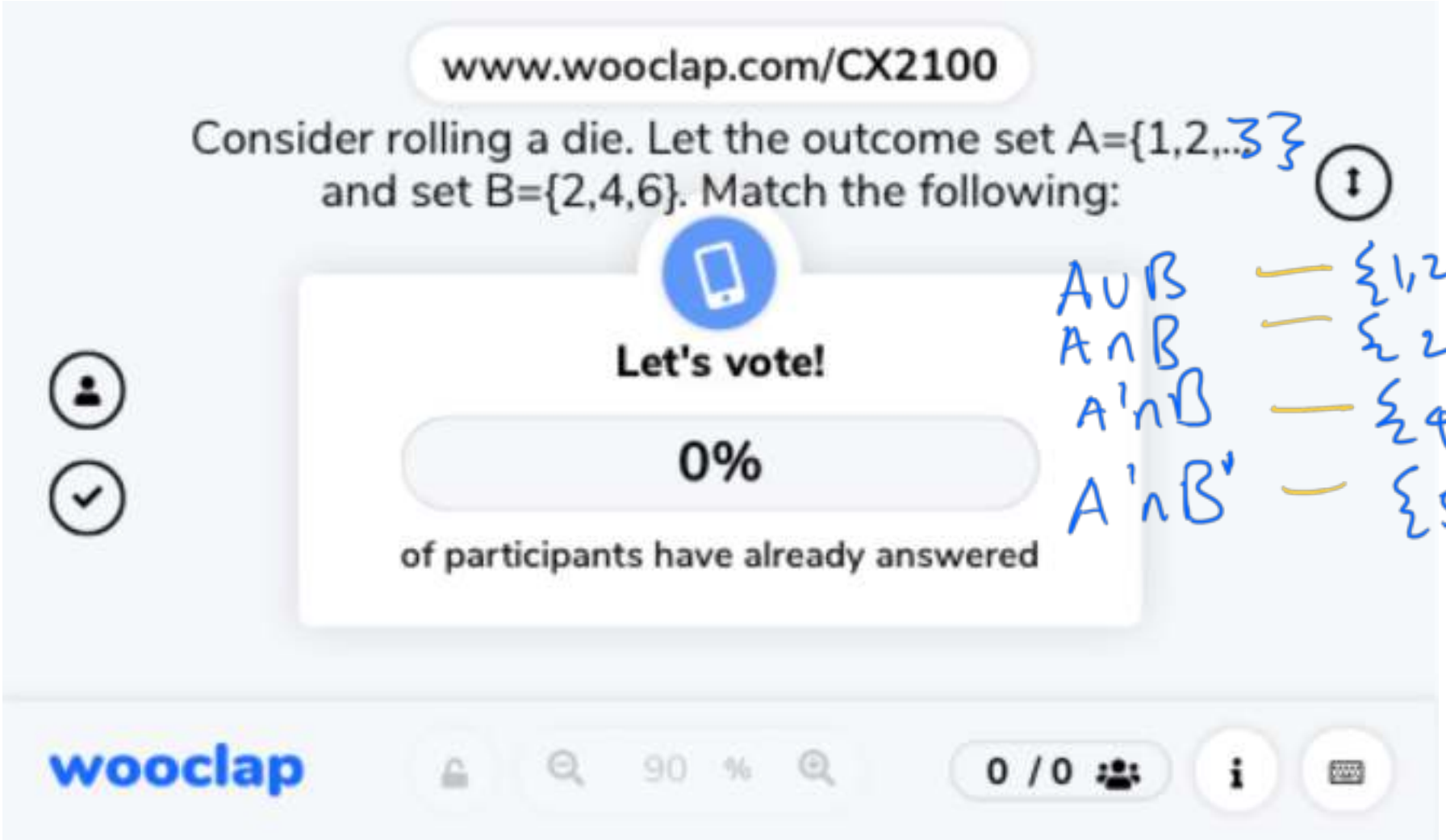


Set A ⊂ Ω. Complement of A, denoted by A′, consists of all elements in Ω that are not in A.
Eg: Ω = {1,2,3,4,5,6}, A={1,2}, A′={3,4,5,6}

Union of A and B consists of all elements in A or B. Notation: A ∪ B.
Eg: {1,2,3} ∪ {2,3,4} = {1,2,3,4}

Intersection of A and B consists of all elements in both A and B. Notation: A ∩ B.
Eg: {1,2,3} ∩ {2,3,4} = {2,3}

Consider rolling a die. Let the outcome set A={1,2,3} and set B={2,4,6}. Match the following:



- Counting

When solving probability problems, we may involve counting.

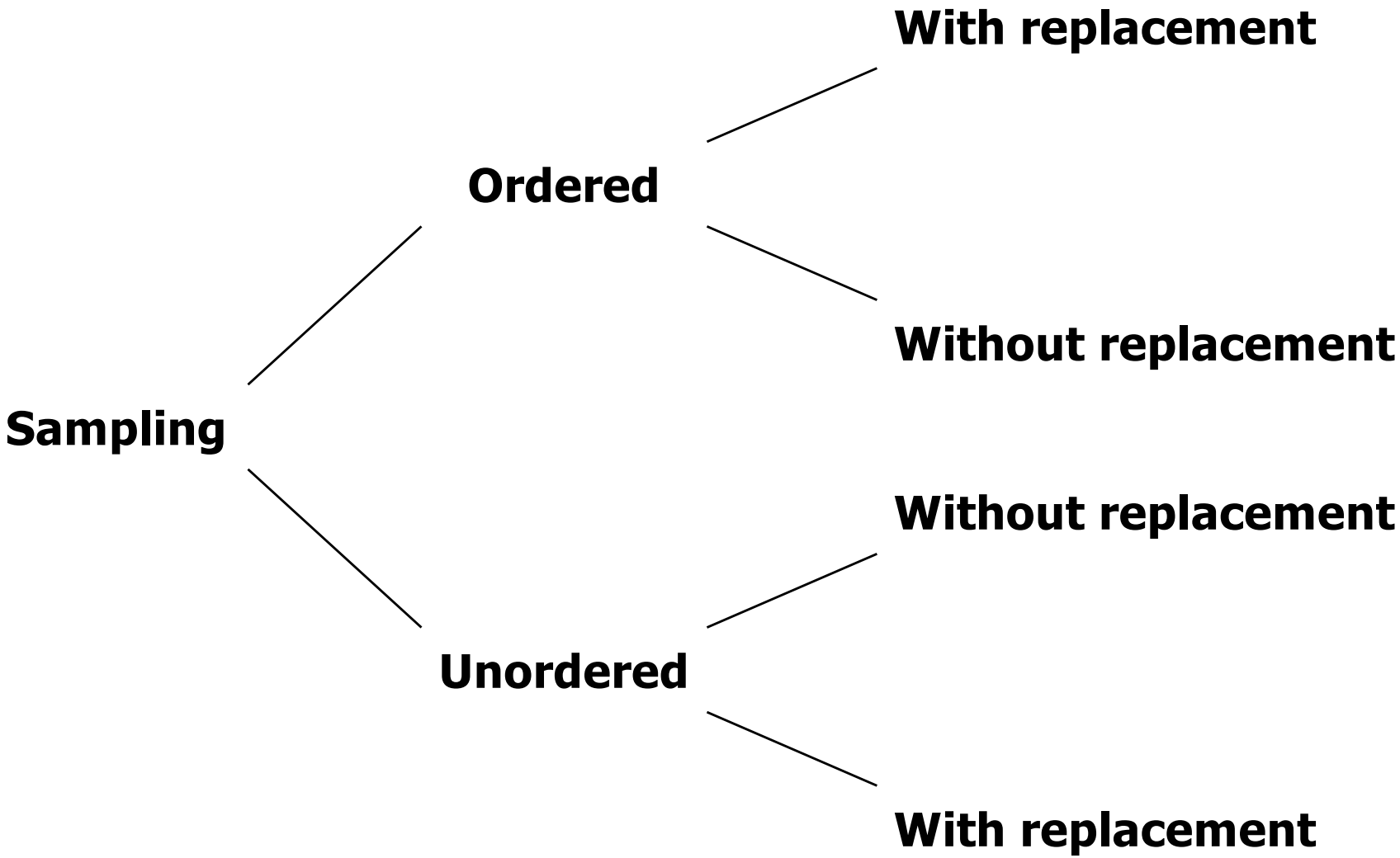
Eg: Calculating the probability of guessing the correct 4 digits numeric code.

Each digit is sampled from {0,1,2,…,9}.

Different scenarios for consideration:

- sampling with replacement
- sampling without replacement
- digits drawn are ordered
- digits drawn are unordered

- Counting – general scenario



- Counting – ordered sampling with replacement

  Ordered sampling k elements from a set of n elements with replacement.

  No. of possible arrangements = n^k

  Eg: Calculate the total number of different 4-digit numeric codes. Each digit is independently selected from {0,1,2,…,9}.

  1st digit: 10 possible numbers

  2nd digit: also 10 possible numbers.

  3rd and 4th digit: each also 10 possible numbers.

  Total arrangements = 10 x 10 x 10 x 10 = 10^4

- Counting – ordered sampling without replacement

Ordered sampling k elements from a set of n elements without replacement.

No. of possible arrangements =

n x $(n-1)$ x $(n-2)$ x $(n-3)$ x ... x $(n-k+1)$

Number of k permutations of n elements:

$$^{n}\mathrm{P}_{k} = \frac{n!}{(n-k)!}$$

$n = 10$ $k = 4$

Eg: Calculate the total number of different 4-digit numeric codes. Each digit is sampled from {0,1,2,…,9} without replacement.

1st digit: 10 possible numbers

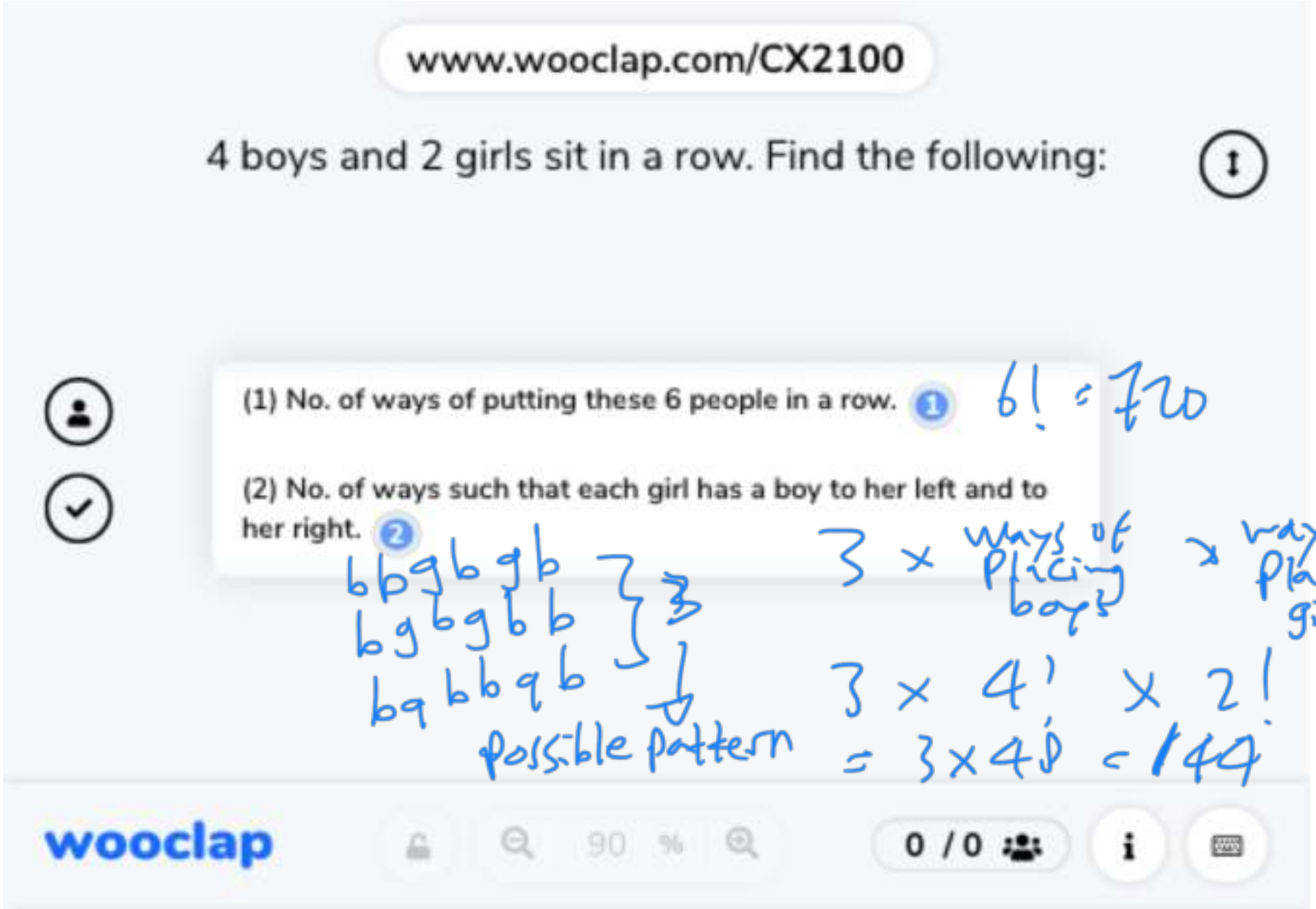
2nd digit: 9 possible numbers

3rd digit: 8 possible numbers and so on.

Total arrangements = 10 x 9 x 8 x 7

i.e. there are ${}^{10}P_4 = \frac{n!}{(n-k)!} = \frac{10!}{(10-4)!}$ arrangements

Eg: 4 boys and 2 girls sit in a row. Find the following:



- Counting – unordered sampling without replacement

We choose k elements from a set of n elements and the ordering does not matter.

Number of arrangements equals to k combinations of n elements

$$\binom{n}{n-k} = \binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

use for binomial too

Note: $\binom{n}{k} = \frac{n!}{k!\,(n-k)!} = \binom{n}{n-k}$

Eg: How many combinations of two numbers between 1 and 6 are there:

Ans: $\frac{6!}{2!\,(6-2)!} = \frac{6 \times 5}{2 \times 1} = 15$

- Counting – unordered sampling with replacement

Pick k elements from a set of n elements, one at a time with replacement and the ordering does not matter.

Eg: Pick two numbers from {1, 2, 3}, we have

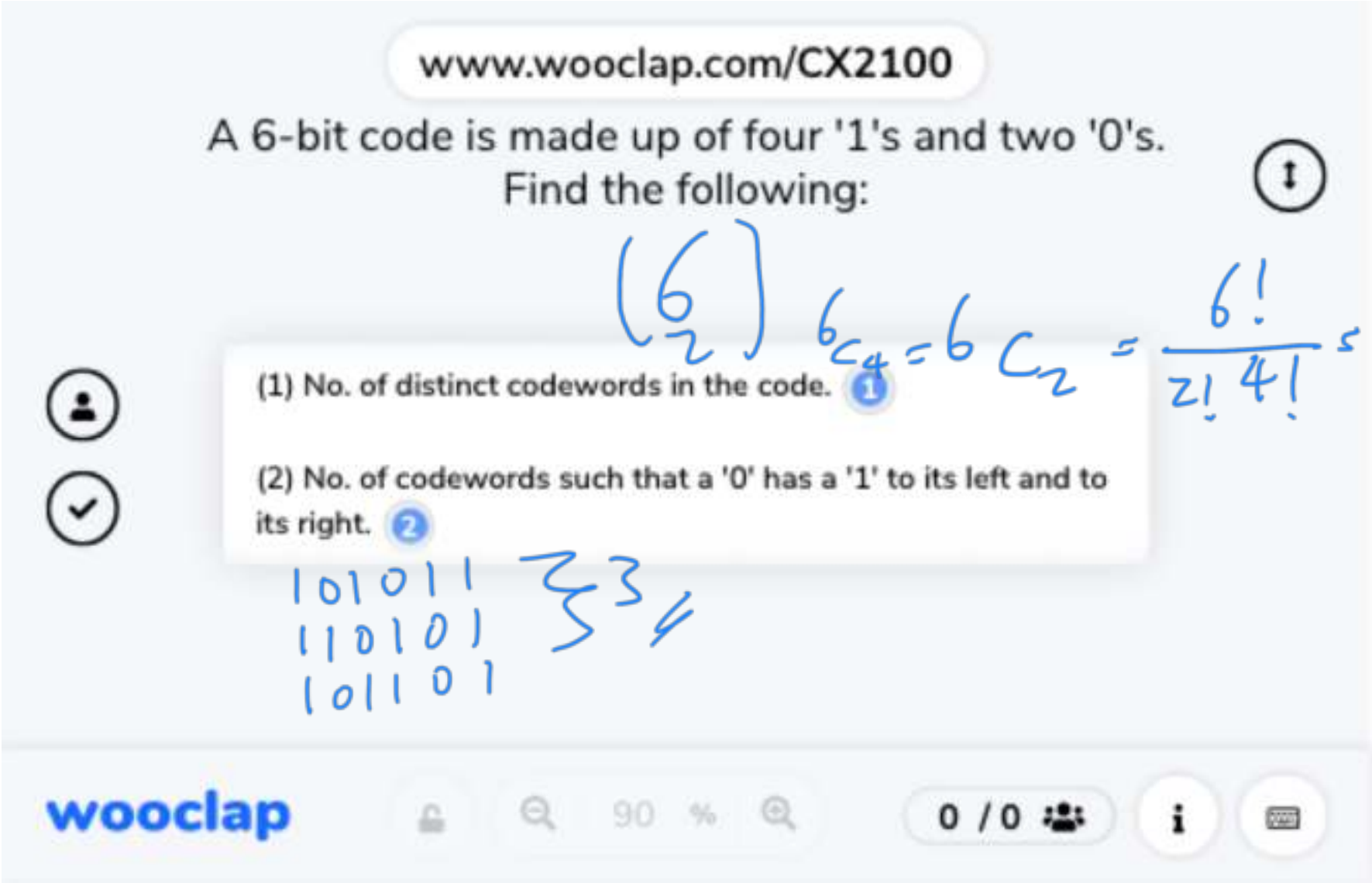
6 arrangements with n=3 and k=2.

i.e. (1,1), (1,2), (1,3), (2,2), (2,3) and (3,3)

In general:

No. of arrangements = $\binom{n+k-1}{k}$

Eg: A 6-bit code is made up of 4 1's and 2 0's.
Calculate:



- Probability Theory

In a random experiment, the outcome of the experiment occurs with a certain probability.

Eg: We toss a coin. If the coin is unbiased, then the chances of getting a Head is 50%.

Eg: We roll a fair die two times. The outcome $\in \{11, 12, 13, 14, 15, 16, 21, 22, \ldots, 65, 66\}$.

Total of 36 possible outcomes.

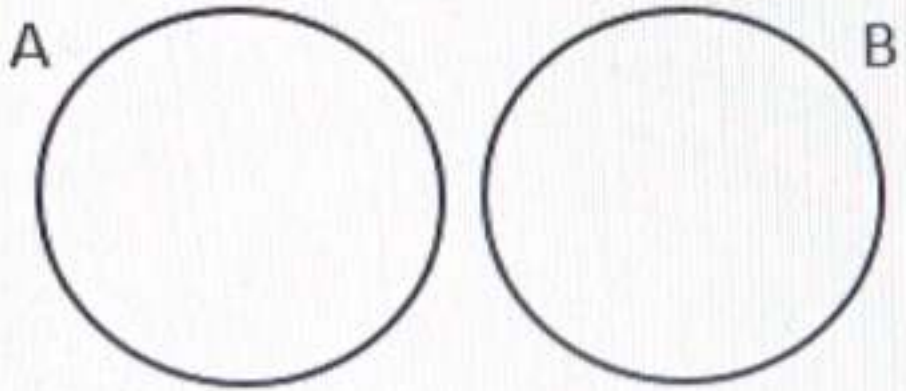
Probability of getting a "1" followed by another "1" is 1/36.

Probability of getting 1^{st} no. = 2^{nd} no. is 6/36.

Definitions

- A sample space is the set S of all possible outcomes of an experiment.
- An event is a set of one or more (favorable) outcomes in the sample space.
- Two events are mutually exclusive if they have no outcomes in common.
- They are exhaustive if they cover all possible outcomes.
- Two events are independent if the probability that one occurs is not affected by whether or not the other has occurred.

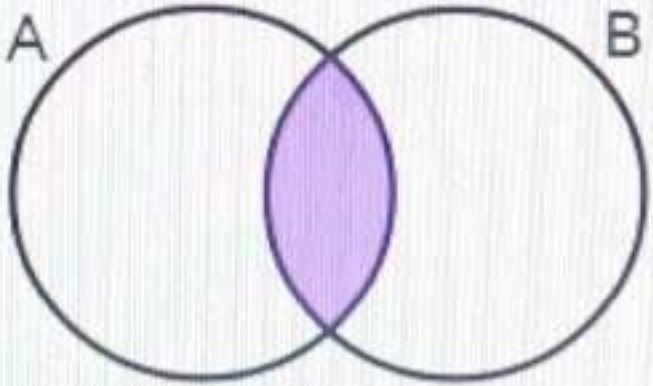
Mutually Exclusive Events



$P(A \text{ or } B) = P(A) + P(B)$

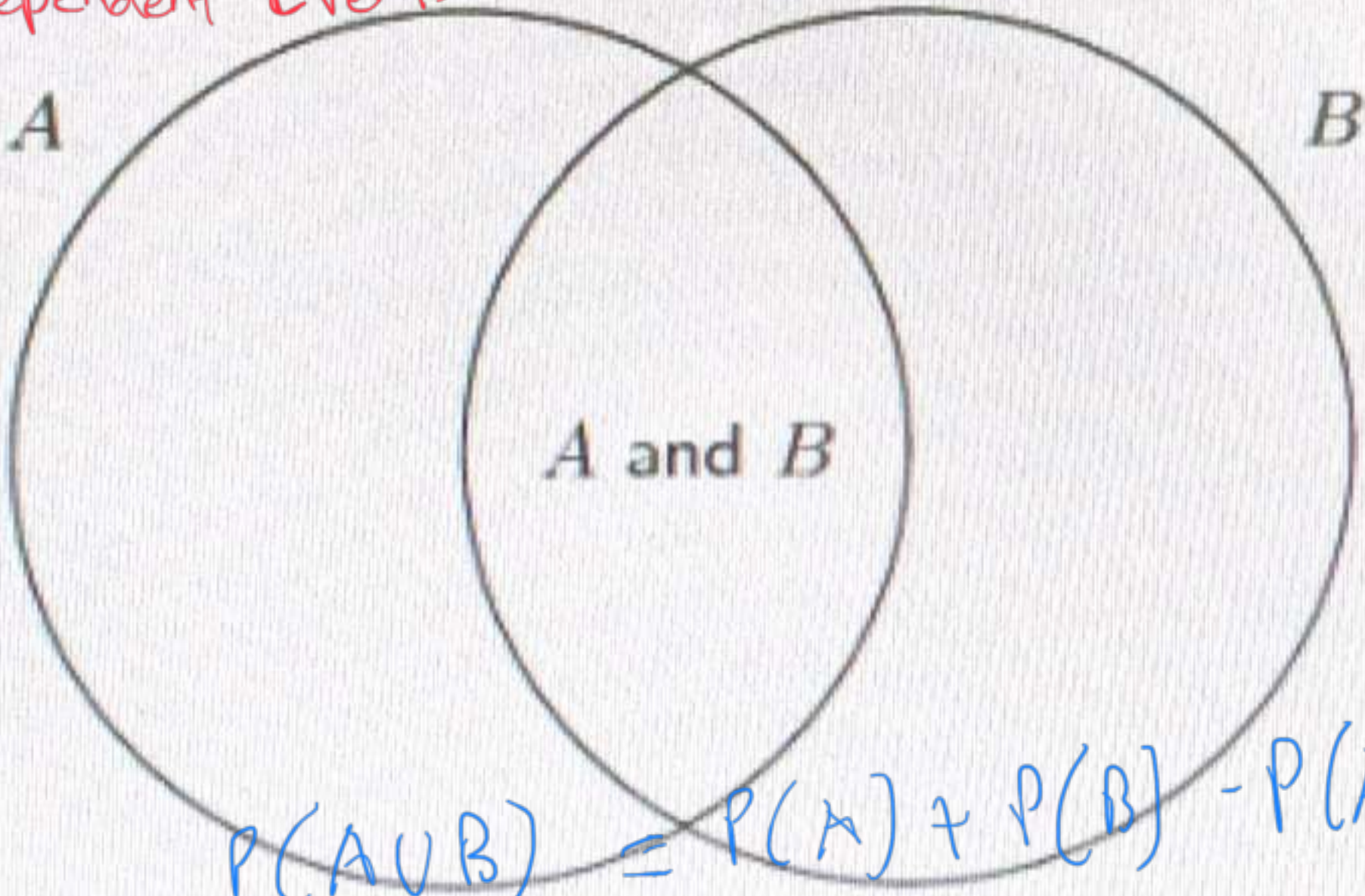
$P(A \cap B) = 0$

Non-Mutually Exclusive Events



$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

Independent Events



$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A \cap B) = P(A) \cdot P(B)$$

Axioms of Probability:

(1) Probability of any outcome or event X is a non-negative:

$$P(X) \geq 0$$

(2) Probability of the sample space S is 1:

$$P(S) = 1$$

(3) If X_1, X_2, X_3, … are mutually exclusive events, then:

$$P(X_1 \text{ or } X_2 \text{ or } X_3 \ldots) = P(X_1) + P(X_2) + P(X_3) + \ldots$$

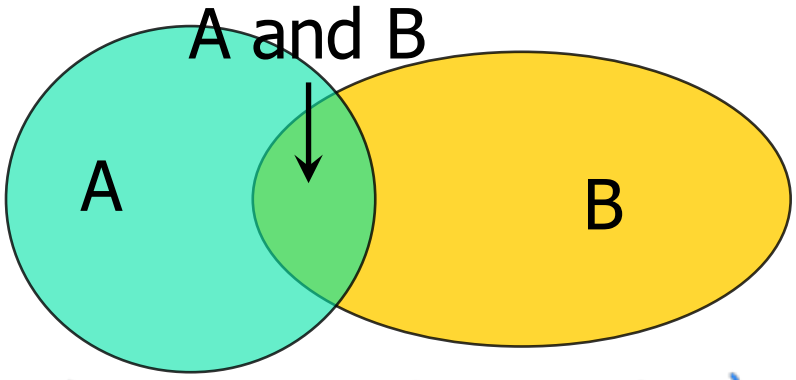
Probabilities and Events (mutually exclusive)

Eg: A company has decided that in the next 5 years, 40% of their new employees will be men, 30% will be Singaporean, and 35% will be foreigner women. What percentage of new employees will be Singaporean men?

	Men	Women	Total
Foreigner	0.35	0.35	0.7
Singaporean	0.05	0.25	0.3
Total	0.4	0.6	1

Probabilities and Events (non-mutually exclusive)

Eg: A certain kind of fruit is grown in 2 districts, A and B. Both areas sometimes get fruitflies. Suppose the probabilities are P(A)=0.1, P(B)=0.05 and P(A and B)=0.02, what is the probability that one or other (or both) districts are infected at a given time?



$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
$$= 0.1 + 0.05 - 0.02$$
$$= 0.13$$

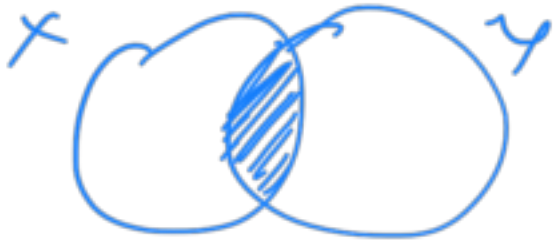
Probabilities and Events (independent)

Eg: A company has 2 guards. Each carries a pager activated by sensors. Guard 1 and guard 2 respond to pager alert 80% and 50% of the time respectively. They independently report any alert. What is the probability that at least one will report an alert?

$X = G1 = 0.8 = P(X)$

$Y = G2 = 0.5 = P(Y)$

2 events



probability of 2 independent event $= P(X) \cdot P(Y)$

$= 0.4$

$P(\text{at least one reports}) = P(X) + P(Y) - P(X \cap Y)$

$= 0.8 + 0.5 - 0.4$

$= 0.9$ //

Conditional Probabilities

$P(X \mid Y)$
Lo given that

Assuming that 10 percent of the days are rainy in a certain city.

$$P(\text{rain}) = 0.1$$

The probability that it rains given that it is cloudy might be, say,

$$P(\text{rain}|\text{cloudy}) = 0.8$$

This is known as conditional probability: the probability of event A given that event B has occurred.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \text{where } B > 0$$

Conditional Probabilities

www.wooclap.com/CX2100

Fig 1.

Fig 2.

Given P(A)=0.5 and P(B)=0.8, determine the following.
(1) P(A|B) for Fig 1.
(2) P(A|B) for Fig 2.
(3) P(B|A) for Fig 2.

The most frequent answers are

(1)

wooclap

90 %

0 / 0

Conditional Probabilities

Eg: A jar contains 10 blue, 5 red, 4 green and 1 yellow marbles. Two marbles are randomly picked. What is the probability that one will be blue and the other yellow?

$P(1^{st} B \cap 2^{nd} Y) = P(1^{st} B) \cdot P(2^{nd} Y \mid 1^{st} B)$

$= 10/20 \cdot 1/19 = 0.0263$

$P(1^{st} Y \cap 2^{nd} B) = P(1^{st} Y) \cdot P(2^{nd} B \mid 1^{st} Y)$

$= 1/20 \cdot 10/19 = 0.0263$

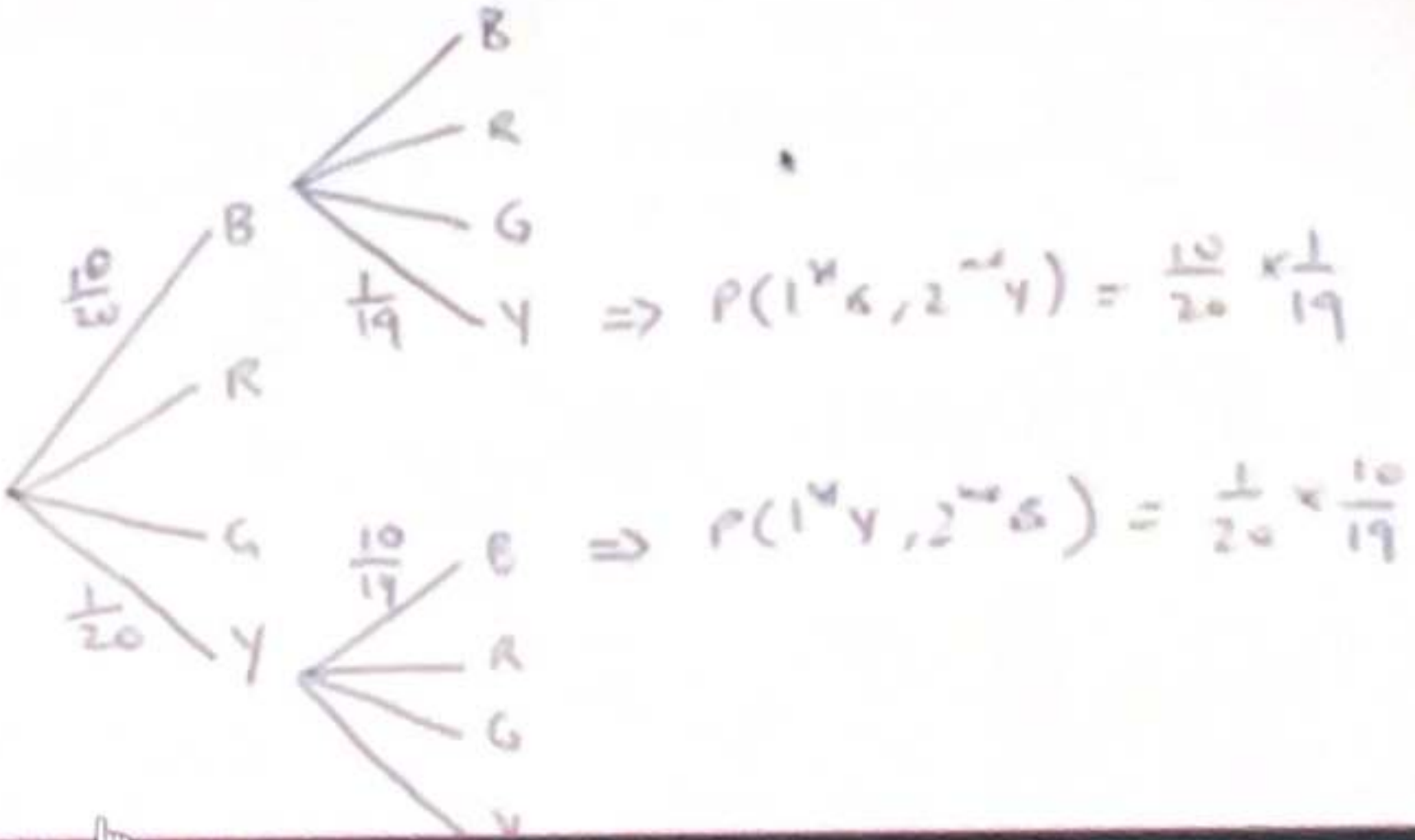
$P(1 \text{ blue} \cap 1 \text{ yellow}) = P(1^{st} B \cap 2^{nd} Y) + P(1^{st} Y \cap 2^{nd} B)$

$= 0.0526$

Illustration using a tree diagram:

Eg: A jar contains 10 blue, 5 red, 4 green and 1 yellow marbles. Probability of getting one blue and one yellow:

Illustration using a tree diagram:

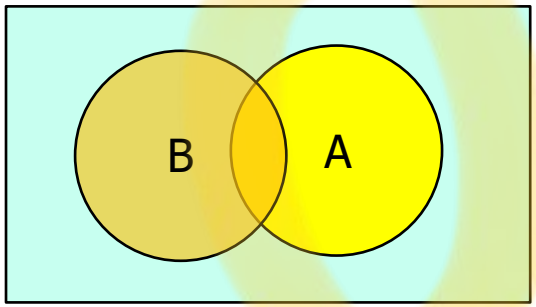


Bayes' Theorem

Given two events A and B, where P(A) > 0, we have:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} = \frac{P(A \cap B)}{P(A)}$$

Note: $P(A) = P(A \cap B) + P(A \cap B')$



Similarly:

$$P(A|B) = \frac{P(B|A)P(B)}{P(B)} = \frac{P(A \cap B)}{P(B)}$$

Bayes' Theorem

Eg: A test correctly identifies a disease in 95% of people who have it. It correctly identifies no disease in 94% of people who do not have it. In the population, 3% of the people have the disease. What is the probability that one has the disease if tested positive?

$P(D) = 0.03 \rightarrow P(D') = 0.97$ D = one has disease

$P(T'|D') = 0.94 \rightarrow P(T|D') = 0.06$ T = test is +

$P(T|D) = 0.95 \rightarrow P(T'|D) = 0.05$

$$P(D|T) = \frac{P(D \cap T)}{P(T)} = \frac{P(T|D)\,P(D)}{P(T \cap D) + P(T \cap D')} = \frac{0.95 \cdot 0.03}{P(T|D)P(D) + P(T|D')P(D')}$$

$$= \frac{0.0285}{0.95 \cdot 0.03 + 0.06 \cdot 0.97} = \frac{0.0285}{0.0867} = 0.329$$

Summary For Probability Calculations

Are X and Y mutually exclusive?

Yes

1. P(X and Y) = 0 because event (X and Y) is impossible
2. P(X or Y) = P(X) + P(Y)

No

1. P(X and Y) > 0
2. P(X or Y) = P(X) + P(Y) – P(X and Y)

Are X and Y independent?

Yes

1. P(X|Y) = P(X)
2. P(X and Y) = P(X) P(Y)

No

1. $P(X|Y) = \frac{P(X \text{ and } Y)}{P(Y)}$
2. P(X and Y) = P(X|Y) P(Y)
   = P(Y|X) P(X)

Ch 6. Probability Distributions

- Random variables – Discrete
  – Continuous
- Discrete Probability distributions – Binomial, Poisson, Geometric
- Continuous Probability distributions – Uniform, Exponential, Normal

Random Variables

The outcomes of random experiments commonly takes on numerical values.

If the outcomes are not numerical, they can be represented in terms of numbers (this facilitates mathematical analysis).

A random variable (r.v.) is a variable that has a numerical value which is defined on or determined by the outcomes or events of an experiment. The numerical value cannot be predicted exactly but can be described by its probability.

Random Variables

Random variables can be discrete or continuous.

Eg: The no. of seeds germinating from a flower pot. Possible values for X are 0, 1, 2, … (discrete).

The maximum daily temperature in Singapore. Possible values are 0 – 50°C, e.g. 26.1276°C (continuous).

The response to questions with "Yes", "No", or "Don't know" answers. This are not a r.v. (the values are not numerical).

Let Y be the number of "Yes"s.
Y is a discrete r.v.

Random Variables

Expected Value of a r.v.

For a given set of observed data $\{x_1, x_2, \ldots, x_n\}$, one of the central tendency measure is the arithmetic average:

$$\bar{X} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

$\bar{X}$ is calculated based on a sample of n observed values.

If we know the probability of occurring associated with each value of a r.v., then we can obtained the weighted average which is known as the expectation of the r.v.

Random Variables

Expected Value of a r.v.

The probability distribution or probability mass function (pmf) of a r.v. X is given as:

Values of X	x_1	x_2	...	x_k
Probabilities	p_1	p_2	...	p_k

Its expected value is defined as:

$$E(X) = x_1 p_1 + x_2 p_2 + \ldots + x_k p_k = \sum_{i=1}^{k} x_i p_i$$

Also denoted by μ_X

Random Variables

Eg: (Decision Analysis)

An oil exploration company has a lease for which it must decide on one of the following options:

1. Sell now and can get $125K.
2. Hold for a year and if oil prices rise (prob = 0.6) it can sell for $300K or if oil prices fall (prob = 0.4) it can get $100K. $0.6 \times 300K + 0.4 \times 100K = 180K + 40K = \$220K$
3. Drill now. The cost of drilling is $200K and drilling will lead to one of the following outcomes: $-100K + 80K + 130K = \$110K$

Well type	Dry	Wet	Gusher
Probability	0.5	0.4	0.1
Profit	$0	$400K	$1500K

-200 200 1300

What should the company do? (2)

Random Variables

Solution: Let X be the financial gain.

Option 1

Option 2
(hold for 1 yr):

X	300	100
Probability	0.6	0.4

$\therefore E(X) =$

Option 3, X = profit – drilling costs ($200K)

X	0 – 200K	400K – 200K	1,500K – 200K
Probability	0.5	0.4	0.1

$\therefore E(X) =$

Decision:

Wooclap Ex:
Given the pmf of X below, if $Y=(X-1)^2$, find E[Y].

x	0	1	2
$P_X(x)$	0.2	0.5	0.3

Random Variables

Variance of a r.v.

Recall that variance is a measure of variability. For a sample of n independent observations the sample variance is obtained using:

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{X})^2}{n-1} \quad \text{or} \quad \frac{1}{n-1}\left[\sum x^2 - \frac{(\sum x)^2}{n}\right]$$

The variance of a r.v. X is defined as

$$\begin{aligned} \text{var}(X) &= p_1(x_1-\mu)^2 + p_2(x_2-\mu)^2 + \cdots + p_n(x_n-\mu)^2 \\ &= \sum_{i=1}^{n} p_i(x_i-\mu)^2 = E\left[(x-\mu)^2\right] \end{aligned}$$

where p_i = probability that x_i occurs.

Random Variables

It represents the theoretical limit of the sample variance s^2 as the sample size n becomes very large. Var(X) is often denoted by σ^2.

An alternate formula for var(X) is as follows:

$$\text{var}(X) = \Sigma\, p_i (x_i^2 - 2x_i\mu + \mu^2)$$

$$= (p_1 x_1^2 + \ldots + p_n x_n^2) - \mu^2$$

$$= E[X^2] - \mu^2 \quad \text{or} \quad E[X^2] - (E[X])^2$$

Random Variables

Eg

Let the random variable X be the number of male students in a group from a class of size 5. The probability distribution of X is

X	0	1	2	3	4	5
$P(X = x)$	1/32	5/32	10/32	10/32	5/32	1/32

What is the expected number of males $E(X)$ in the group, and what is the standard deviation, σ, of X?

Random Variables

Solution: $E(X) = \sum_{i=1}^{n} x_i P(x_i)$

$\mu = 0 + \frac{5}{32} + \frac{20}{32} + \frac{30}{32} + \frac{20}{30} + \frac{5}{32}$

$\mu = \frac{80}{32} = 2.5$

$\text{var}(X) = \sum x^2 P(x) - \mu^2$

$\left(0^2 \cdot \frac{1}{32} + 1^2 \frac{5}{32} + 2^2 \frac{10}{32} + 3^2 \frac{10}{32} + 4^2 \frac{5}{32} + 5^2 \frac{1}{32}\right) - M^2$

$\left(0 + \frac{5}{32} + \frac{40}{32} + \frac{90}{32} + \frac{80}{32} + \frac{25}{32}\right) - (2.5)^2$

$\sigma^2 = 7.5 - 6.25 = 1.25$

$\sigma = 1.12$

Wooclap Ex:
Given the pmf of X below, if $Y=(X-1)^2$, find Var[Y].

x	0	1	2
$P_X(x)$	0.2	0.5	0.3

$Var[Y] = E[Y^2] - E[Y]^2$

$E[Y^2] = 1 \times 0.2 + 0 \times 0.5 + 1 \times 0.3 = 0.5$

$E[Y] = 0.5$

$Var[Y] = 0.5 - 0.5^2 = 0.25$

Random Variables

Empirical quantity (m observed data)	Theoretical quantity (mathematical d.r.v.)	Remarks
Relative freq of x_i is $\frac{f_i}{m}$	$P[X = x_i] = p_i$	$\frac{f_i}{n} \to p_i$ as $n \to \infty$
$\sum_{i=1}^{m} \frac{f_i}{m} = 1$	$\sum_{i=1}^{n} p_i = 1$	
Mean: $\bar{X} = \sum_{i=1}^{m} \frac{f_i}{m} x_i$	Expectation: $E[X] = \mu = \sum_{i=1}^{n} p_i \, x_i$	$\bar{X} \to \mu$ as $n \to \infty$
Variance: $s^2 = \sum_{i=1}^{m} \frac{(x_i - \bar{X})^2 f_i}{m-1}$	Var[X]: $\sigma^2 = \sum_{i=1}^{n} (x_i - \bar{X})^2 p_i$	$s^2 \to \sigma^2$ as $n \to \infty$

Random Variables

Expected Value and Variance for a Function of r.v.

(Linear transformation of r.v.)

If $Y = a + bX$ where X is a r.v. and a and b are known constant values, then

$$E(Y) = a + bE(X)$$

and

$$\text{var}(Y) = b^2 \text{var}(X)$$

so

$$\sigma_y = \sqrt{b^2 \text{var}(X)} = \sqrt{b^2 \sigma_x^2} = b\sigma_x$$

Random Variables

Similarly if $T = a + bX + cY$ where X and Y are uncorrelated r.v. and a, b and c are known constants, then

$$E(T) = a + bE(X) + cE(Y)$$

and

$$\text{var}(T) = b^2 \text{var}(X) + c^2 \text{var}(Y)$$

Random Variables

Eg: A company makes products for local and export markets. The number of sales for next year are estimated as follows:

local, X units	1,000	3,000	5,000	10,000
probability	0.1	0.3	0.4	0.2

export, Y units	300	500	700
probability	0.4	0.5	0.1

Hence $E(X) =$ $100 + 900 + 2000 + 2000$
$= 5000$

Random Variables

$E(Y) =$ 120 + 250 + 70 = 440

Suppose the company makes a profit of $2000 on each unit sold on the local market and $3500 on each exported unit.

Hence the total profit is $2000X + 3500Y$

And the expected profit is 2000.5000 + 3500.440

1540000

$E(T) =$

Random Variables

Consider a component is made by cutting a piece of metal to length X and trimming it by amount Y. Both processes are somewhat imprecise.

Nett length $T = bX + cY$ with $b = 1$ and $c = -1$.

$$E(T) = b\,E(X) + c\,E(Y) = E(X) - E(Y)$$

$$\text{Var}(T) = b^2\,\text{var}(X) + c^2\,\text{var}(Y)$$

Note

i.e. var(T) is greater than either var(X) or var(Y), even though $T = X - Y$, because both X and Y contribute to the variability in T.

Prob. Distribution of Discrete r.v.

Special Probability Distributions

- Binomial Distribution
- Poisson Distribution
- Geometric Distribution

Prob. Distribution of Discrete r.v.

The list of possible values that a d.r.v. X can take and their probabilities is called the discrete probability distribution (also known as probability mass function, pmf) for X.

Eg:

Let the random variable X be the number of girls in a family of 3 children.

Possible values:

$X = 3$	*GGG*		
$X = 2$	*GGB*	*GBG*	*BGG*
$X = 1$	*BBG*	*BGB*	*GBB*
$X = 0$	*BBB*		

Prob. Distribution of Discrete r.v.

Assume that the 8 outcomes are equally likely,

x_i	0	1	2	3
Probability Distribution P($X=x_i$)	1/8	3/8	3/8	1/8
Cumulative Distribution P($X \leq x_i$)	1/8	4/8	7/8	1

In general, we write:

$$P(X = x_i) = p_i \text{ for } i = 1, \ldots, k, \text{ and } 0 \leq p_i \leq 1$$

Notation convention: we often use capital letters for random variables and small letters for specific values.

Prob. Distribution of Discrete r.v.

Eg: Let X denote the sum of the results of 2 dice thrown.

Outcomes:

1,1	**2,1**	**3,1**	**…**	**6,1**
1,2	**2,2**	**3,2**	**…**	**6,2**
⋮	⋮	⋮		⋮
1,6	**2,6**	**3,6**	**…**	**6,6**

36 outcomes and the values of X are 2, 3, …, 12.

Assuming that the outcomes are equally likely, i.e. the probability of each outcome is 1/36, the probability distribution of X is

x	2	3	4	…	10	11	12
$P(X = x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	…	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Exercise: find the cumulative distribution.

$\frac{1}{36}$ $\frac{3}{36}$ $\frac{6}{36}$ - - - $\frac{33}{36}$ $\frac{35}{36}$ 1

Prob. Distribution of Discrete r.v.

Binomial Distribution

- Consider n Bernoulli trials, where each trial is an "experiment" with exactly 2 possible outcomes, "success" and "failure".
- Assume that the probability of success (S) is the same for all trials, $P(S) = p$, and the probability of failure $P(F) = 1 - p$.
- Assume also that trials are independent, i.e., the probability for any given combination of successes and failures can be obtained by multiplying the probabilities for each trial outcome.

Prob. Distribution of Discrete r.v.

Eg: Consider 5 trials, the probability,

$$P(SSFSF) = p \cdot p \cdot (1-p) \cdot p \cdot (1-p)$$

$$= p^3(1-p)^2$$

The probability of obtaining any 3 successes and 2 failures in 5 trials, i.e. *SSSFF*, *SSFSF*, ... etc., is $p^3(1-p)^2$ for each of the different ways this could occur.

Prob. Distribution of Discrete r.v.

The number of distinct "arrangements" of 3 successes and 2 failures can be calculated using the binomial coefficient $\binom{n}{x}$ or ${}^{n}C_{x}$

The binomial coefficient is defined as: $${}^{n}C_{x} = \frac{n!}{x!\ (n-x)!}$$

where $n! = n \times (n-1) \times \ldots \times 2 \times 1$

In this example, $$\binom{5}{3} = \frac{5!}{3!2!} = \frac{5\times4\times3\times2\times1}{(3\times2\times1)(2\times1)} = 10$$

so there are 10 distinct ways to obtain 3 successes in 5 trials, with each arrangement having a probability of $p^3(1-p)^2$.

Prob. Distribution of Discrete r.v.

Let X be the r.v. equal to the total number of successes in n trials. To calculate the probability of obtaining x successes, we have:

$$P(X = x) = \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x}$$

arrangements of x S's and $(n\text{-}x)$ F's

prob. of x S's

prob. of $(n\text{-}x)$ F's

The distribution of the count of successes is called the binomial distribution with two parameters, n and p. We say $X \sim B(n, p)$.

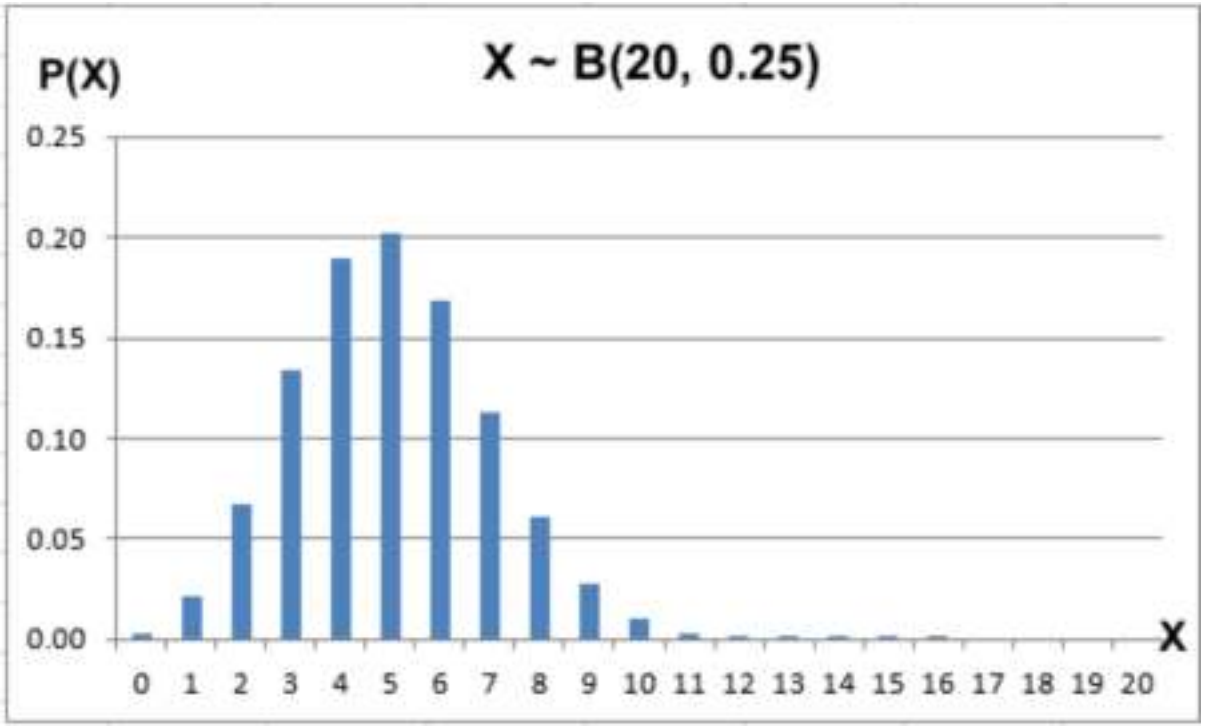
X	0	1	...	n
$P(X=x)$	${}^nC_0 p^0 (1-p)^n$	${}^nC_1 p^1 (1-p)^{n-1}$	...	${}^nC_n p^n (1-p)^{n-n}$

Prob. Distribution of Discrete r.v.

Binomial Distribution – X ~ B(*n*,*p*)

$$P(X = x) = \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x}$$

Example:



Prob. Distribution of Discrete r.v.

Mean and variance of X ~ B(*n*,*p*)

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

The mean of X is given by:

$$E[X] = \sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x} = np$$

And the variance is:

$$Var[X] = E[X^2] - E[X]^2 = np(1-p) = npq$$

where $q = (1-p)$

Prob. Distribution of Discrete r.v.

Example:

A football team plays 3 games. Assume each game is a Bernoulli trial with prob(win) = 0.5. Let the r.v. X be the number of wins. What is the probability that the team will win exactly 2 games?

$$\binom{3}{2} 0.5^2 (0.5)^1 = \frac{3!}{2!1!} \cdot 0.25 \cdot 0.5$$

$$= 0.375$$

Solution

X has binomial distribution with $n = 3$ and $p = 0.5$, outcome Win(W) or Lose(L) on each trial. i.e.

$X \sim B(3, 0.5)$

Prob. Distribution of Discrete r.v.

Using the formula for Binomial probabilities:

$$P(X=x)=\binom{n}{x}p^{x}(1-p)^{n-x}$$

$$= \binom{3}{2} 0.5^{2}(1-0.5)^{3-2}$$

$$= 0375$$

Woocla

There are 10 blue and 30 green balls in a bag. A ball is random picked, the colour is noted and it is put back into the bag. This process is repeated 10 times. Let X be the number of times a blue ball was picked, Calculate E[X] and Var[X].

Go to www.wooclap.com and use the code **CX2100**

There are 10 blue and 30 green balls in a bag. A ball is random picked, the...

$P(x) = \frac{10}{40} = 0.25$

$E[x] = np = 10 \cdot 0.25 = 2.5$

The expected value of X is

The variance of X is

$Var[x] = npq = 10 \cdot 0.25 \, 0.75 = 1.875$

wooclap

Prob. Distribution of Discrete r.v.

Example:

A quality control system selects a sample of 10 items from each batch of products for testing. If 2 or more of the items are defective the whole batch is rejected.

If the probability of an item being defective is 0.05, what is the probability of

(i) having 2 defectives in the sample?

(ii) the batch being rejected?

Prob. Distribution of Discrete r.v.

Soln: Let X be the number of defectives in the sample of $n = 10$ items. $X \sim$ B (10, 0.05)

(i) $P(x=2) = \binom{10}{2} (0.05)^2 (0.95)^8 = 0.0746$

(ii) $P(X \geq 2) = 1 - P(X=1) - P(X=0)$

$= 1 - \binom{10}{1} (0.05)(0.95)^9 - \binom{10}{0} (0.05)^0 (0.95)^{10}$

$= 1 - 0.3151 - 1 \cdot 1 \cdot 0.5987$

$= 0.0861$

Prob. Distribution of Discrete r.v.

Poisson Distribution – Notation: X~Pois(μ)

- Let r.v. X be the number of "successes" in a given time interval where X is a non-negative integer
- The r.v. X is a Poisson r.v. with probability distribution given by

$$P(X = x) = \frac{e^{-\mu}\mu^{x}}{x!}$$

where μ is the average number of successes and e is a constant = 2.71828...

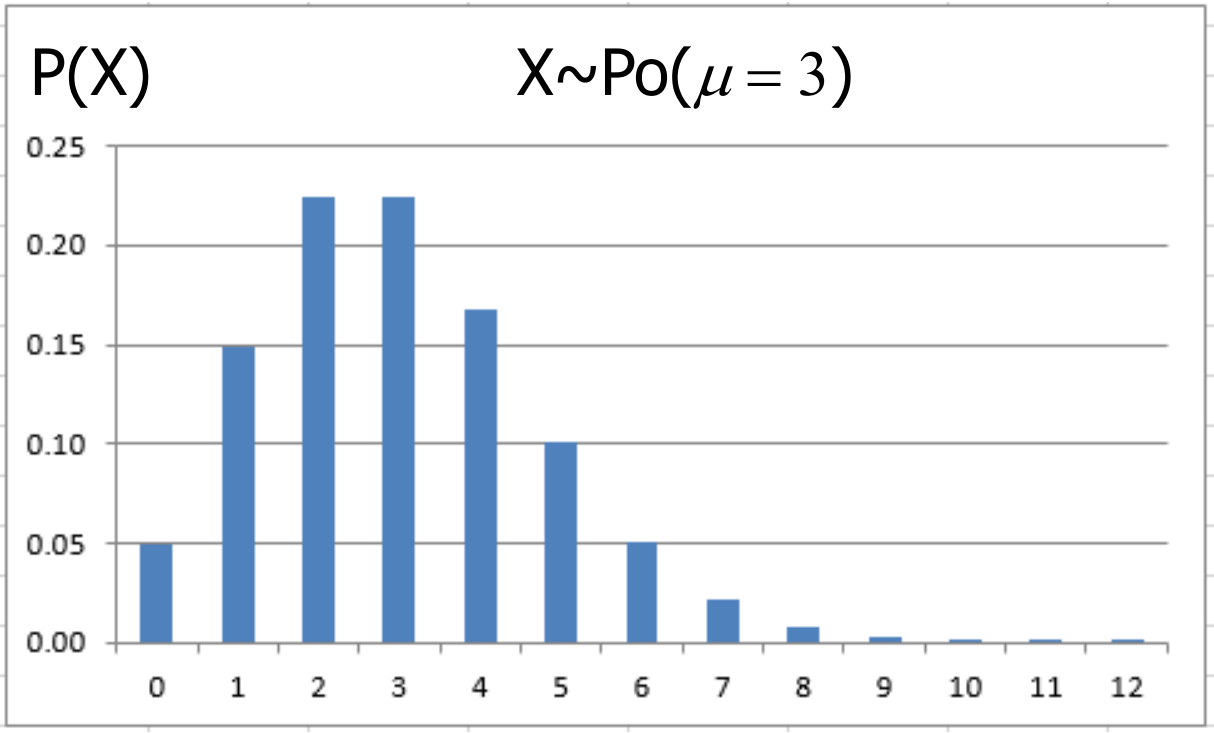
- The expected value and the variance of X are both equal to μ.

Prob. Distribution of Discrete r.v.

Poisson Distribution – X~Pois(μ)

$$P(X = x) = \frac{e^{-\mu}\mu^{x}}{x!}$$

Example:



Prob. Distribution of Discrete r.v.

Eg:

A salesman sells on the average 3 iphones per day. The number of iphones sold per day is a Poisson r.v.

(i) Calculate the probability that in a given day he will sell some iphones.

(ii) Given that there are 8 working hours per day, what is the probability that in a given hour he will sell one iphone?

Prob. Distribution of Discrete r.v.

Soln:

(i) $P(X \geq 1) = 3$ $P(X=0)$

$P(\text{iPhone}) = 1 - P(\text{no iPhone})$

$P(X) \quad 1 - \frac{e^{-3} 3^0}{0!} = 1 - 0.05 = 0.95$

(ii) Avg iphone sold/hour $= 3/8 = 0.375$

$P(X=1) = \frac{e^{-0.375} 0.375^1}{1!} = 0.258$

Eg: A shop selling a certain product of which the weekly demand is a Poisson variable with mean 3. On 1st day of each month, the stocks are replenished. Obtain the minimum number of the product in stock on 1st day of a month so that the shop is at least 50% sure of being able to meet the demands for the month.

X = no of product

$X \sim \text{Pois}(3 \times 4)$ ← monthly

Find largest x that $P(X \leq x) \geq 0.5$

Choose $x = 12$ ← AVG (3×4)

$P(X \leq 12) = 0.576$ $\quad x = 12$ ✓

check $P(X \leq 11) = 0.416$ ✗

Prob. Distribution of Discrete r.v.

Poisson Distribution – Approx to Binomial Dist

For Binomial Distribution $X \sim$ B(n, p) with large n :
i.e. $n \to \infty,\ p \to 0$ and $np \to$ a constant μ

Then it can be shown that :

$$\frac{n!}{x!\,(n-x)!}\,p^x(1-p)^{n-x} \to \frac{e^{-\mu}\mu^x}{x!}$$

$\therefore X \sim$ B(n,p) $\approx$ Pois(μ), where $\mu = np$, we have:

$$P(X = x) = \frac{e^{-\mu}\mu^x}{x!}$$

Poisson Distribution - Approximation to Binomial Distribution

$X \sim B(n, p)$, $\mu = np$

For large n:

$$\lim_{n \to \infty} \binom{n}{x} p^x (1-p)^{n-x} \quad \text{and} \quad \mu = np \Rightarrow p = \frac{\mu}{n}$$

$$= \lim_{n \to \infty} \frac{n!}{(n-x)!\, x!} \left(\frac{\mu}{n}\right)^x \left(1 - \frac{\mu}{n}\right)^{n-x}$$

$$= \lim_{n \to \infty} \frac{n!}{(n-x)!} \frac{1}{n^x} \frac{\mu^x}{x!} \left(1 - \frac{\mu}{n}\right)^n \left(1 - \frac{\mu}{n}\right)$$

$\left(1 - \frac{\mu}{n}\right) \to 1$ as $n \to \infty$

$\frac{n!}{(n-x)!} \frac{1}{n^x} = \frac{n(n-1)(n-2)\cdots(n-x+1)}{n \cdot n \cdot n \cdots n}$ (x terms)

$\to 1$ as $n \to \infty$

$\lim_{n \to \infty} \left(1 - \frac{\mu}{n}\right)^n = e^{-\mu}$

$$= \frac{\mu^x}{x!} e^{-\mu}$$

i.e. $X \sim Po(\mu)$

$\mu = np$

Prob. Distribution of Discrete r.v.

Poisson Distribution – Approx to Binomial Dist

Conditions for which the probability of X~B(n,p) can be approximated by the probability $X \sim \text{Pois}(\mu)$:

- large n (n ≥ 100)
- small p (p ≤ 0.01)
- constant $\mu = np$ ($\mu \leq 20$)

Note that when $p \to 0$, then $(1-p) \to 1$.

∴ the expected value and the variance of X ~B(n,p) are approximately equal.

Intuitive Explanation:

To obtain the mean and variance of the Poisson distribution, consider the binomial distribution under the following conditions:

$n \to \infty$, $p \to 0$ and $np \to \text{a constant } \mu$

The mean and variance of the binomial distribution are given by:

$$E[X] = np$$

$$Var[X] = npq$$

Since $q = 1 - p$, and as $p \to 0$, $q \to 1$ we have:

$E[X] = \mu$ and $Var[X] = npq = \mu$

Example: Suppose in a production line, 1 in 200 LED produced is defective. A random sample of 1000 LEDs are selected. What is the probability that at most 2 of them are defective?

Let X = no of defective LED

$P(1) = 1/200 = 0.005$

$np = 1000 \cdot 0.005 = 5$

$n > 100$, $p < 0.01$ $np < 20$

I use Poisson dist to approx. Binomial dist

$$P(X \le 2) = P(X=0) + P(X=1) + P(X=2)$$

$$= e^{-5} + 5e^{-5} + \frac{5^2 e^{-5}}{2!}$$

$$= 0.125,$$

Prob. Distribution of Discrete r.v.

Geometric Distribution – Notation: X~G(p)

- Consider a sequence of independent Bernoulli trials, where each trial is an "experiment" with exactly 2 possible outcomes, "success" and "failure".
- Assume that the probability of success (S) is the same for all trials, $P(S) = p$, and the probability of failure $P(F) = 1 - p$.
- Each experiment is performed consecutively until the first success is obtained.
- Let X = no. of trials. Hence there are $X - 1$ failures before the 1st success is obtained.

Prob. Distribution of Discrete r.v.

Hence X is a random variable with a Geometric distribution $X \sim G(p)$:

$$\mathrm{P}(X = x) = (1-p)^{x-1}\ p$$

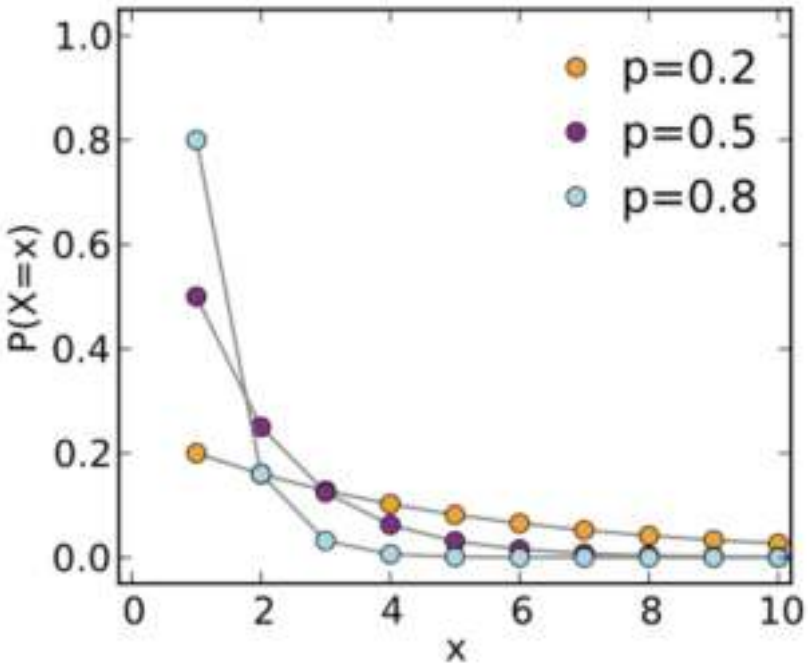
Expected value:

$$\mathrm{E}[\mathrm{X}] = \frac{1}{p}$$

Variance:

$$\mathrm{Var}[\mathrm{X}] = \frac{1-p}{p^2}$$

Examples:



Mean and Variance of Geometric Distribution $X \sim G(p)$

$$E[x] = \sum_{x=1}^{\infty} x (1-p)^{x-1} p$$

this term is $\frac{dq^x}{dq} = x q^{x-1}$

let $q = 1-p$, $\therefore E[x] = p \sum_{x=1}^{\infty} x q^{x-1}$

$= p \frac{d}{dq} \sum_{x=0}^{\infty} q^x$ ← infinite geometric series $= \frac{1}{1-q}$

(1^{st} term $x=0 \Rightarrow 0$)

$= p \frac{d}{dq} (1-q)^{-1} = \frac{p}{(1-q)^2}$

$= \frac{1}{p}$ since $1-q = p$

$Var[x] = E[x^2] - E[x]^2 = E[x(x-1)] + E(x) - E[x]^2$

$E[x(x-1)] = \sum_{x=1}^{\infty} x(x-1) p (1-p)^{x-1}$

$\frac{d}{dq} (x-1) q^{x-2}$

$$Var[X] = E[X^2] - E[X]^2 = E[X(X-1)] + E[X] - E[X]^2$$

$$E[X(X-1)] = \sum_{x=1}^{\infty} x(x-1)\, p\,(1-p)^{x-1}$$

Similar to the above approach, using $\frac{d}{dq}(x-1)q^x$

we will get $\frac{2(1-p)}{p^2}$

$$\therefore Var[X] = \frac{2(1-p)}{p^2} + \frac{1}{p} + \left(\frac{1}{p}\right)^2 = \frac{1-p}{p^2}$$

In a data transmission, the probability that a data frame is received in error is 0.1. If a data frame is

Example:

In a data transmission, the probability that a data frame is received in error is 0.1. If a data frame is received in error, a retransmission will be requested.

a) What is probability that a data frame will be successfully received in no more than 3 transmissions?

b) Determine the average number of transmission per frame.

a) $P = 1 - 0.1 = 0.9$

$P(X=3) = 0.1^{3-1} \cdot 0.9 = 0.01 \cdot 0.9 = 0.009$

$P(X \leq 3) = P(X=1) + P(X=2) + P(X=3)$

$= 1 \times 0.9 + 0.1 \cdot 0.9 + 0.009$

$= 0.999$

b) $E[X] = 1/P = 1.111$

Prob. Distribution of Continuous R.V.

- **Introduction**

- **Special Probability Distributions**

  - Uniform Distribution
  - Exponential Distribution
  - Normal Distribution

Prob. Distribution of Continuous r.v.

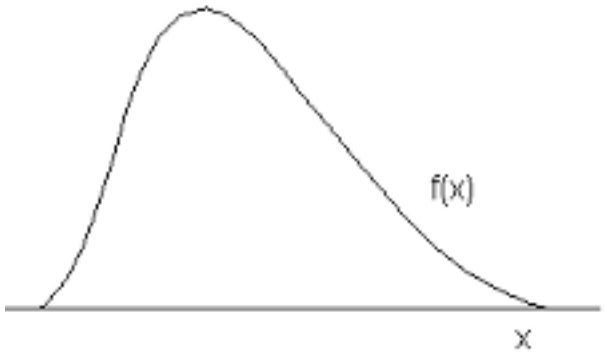
A continuous random variable is a variable used for continuous measurements such as distance, weight, and time.

Probability Density Function (pdf)

For a continuous random variable X, we can describe the probability distribution by some function $f(X)$, such that

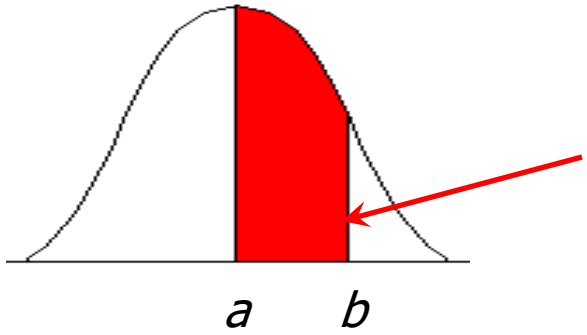
(i) $f(X) \geq 0$ for all X

eg:



Prob. Distribution of Continuous r.v.

(ii) P($a < X < b$) is the area under the curve between a and b :



$$\int_a^b f(x)\,dx$$

(iii) Total area under curve = 1.

The function $f(x)$ is called the probability density function (or pdf) of X.

The probability P($X < x$) is known as the Cumulative distribution function (or CDF) of X.

Wooclap Ex.

Given that the probability density function of X is:

$f(x) = \frac{x^3}{4}$ where $0 < x < c$

Find the value of the constant c.

Given that the probability density function of X is:

$f(x) = \frac{x^3}{4}$ where $0 < x < c$

Find the value of the constant c.

$$\int_0^c \frac{x^3}{4}\,dx = 1$$ Solve for c

$= \frac{x^4}{16}\Big|_0^c = 1$ $c = 2$

Determine the CDF of X.

CDF of X $= \int_{-\infty}^{x} f(x)\ dx$

$$= \int_{-\infty}^{x} \frac{x^3}{4}\ dx = \frac{x^4}{16}\Big|_0^x = \frac{x^4}{16}$$

Prob. Distribution of Continuous r.v.

For a continuous random variable X the probability that X equals to an exact value is zero.

Eg: for the length of a bolt

$$P(X = 1.999965700\,\text{cm}) = 0$$

i.e., $P(X = x) = 0$. Hence, for continuous random variables, probabilities are always associated with a range of values.

Eg: $P(X \leq 2.00\,\text{cm})$ or $P(X > 1.95\,\text{cm})$.

Prob. Distribution of Continuous r.v.

The expected value of a continuous r.v. is defined as:

$$E[X] = \int_{x=-\infty}^{x=\infty} x\, f(x)\, dx = \mu$$

For discrete r.v.

$$E(X) = \sum_{i=1}^{n} p_i x_i$$

and its variance is:

$$\text{var}[X] = \int_{x=-\infty}^{x=\infty} (x-\mu)^2 f(x)\, dx$$

$$= E[X^2] - E[X]^2$$

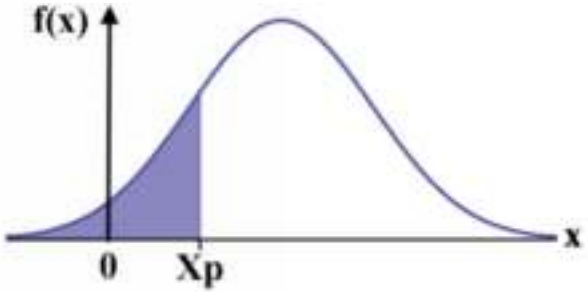
where $$E[X^2] = \int_{x=-\infty}^{x=\infty} x^2 f(x)\, dx$$

For discrete r.v.

$$\text{Var[X]} = \sum_{i=1}^{n} (x_i - \mu)^2 p_i$$

Prob. Distribution of Continuous r.v.

How do we determine the pth percentile from the probability distribution of a continuous r.v.?



The p^{th} percentile is the value X_p such that the probability that $x < X_p$ is p percent.

$$\int_{-\infty}^{X_p} f(x)dx = \frac{p}{100}$$

p^{th} percentile $= X_p$

Wooclap Ex

Given that the probability density function of x is

$$f(x) = e^{-x}, \quad \text{where } x > 0$$

determine the 25th percentile.

$$\int_0^{x_p} e^{-x}\,dx = \frac{25}{100}$$

$$\left[-e^{-x}\right]_0^{x_p} = 0.25$$

$$(-e^{x_p}) - (-e^0) = 0.25$$

$$1 - e^{-x_p} = 0.25$$

$$e^{-x_p} = 0.75$$

$$\ln e^{-x_p} = \ln 0.75$$

$$-x_p = \frac{\ln 0.75}{\ln e} = \frac{\ln(0.75)}{1}$$

$$x_p = 0.288$$

Prob. Distribution of Continuous r.v.

Uniform Distribution

One example of a pdf for a continuous r.v. is the uniform continuous distribution.

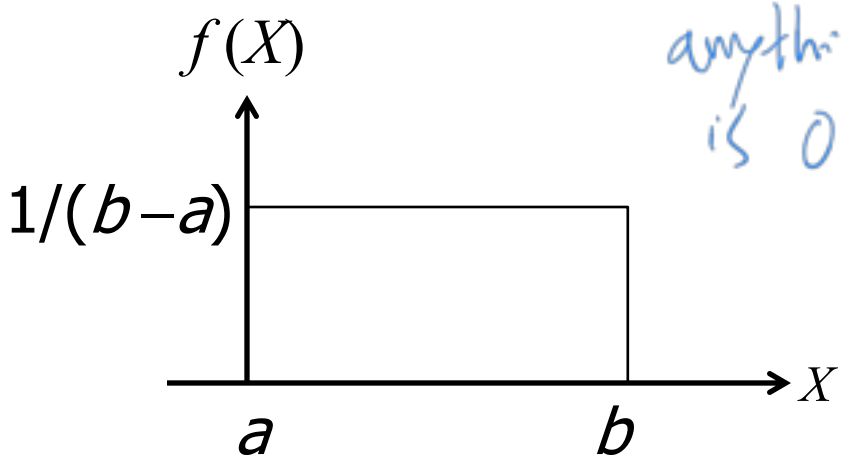
X can take any real value between a and b with uniform probability over this interval.

Notation: $X \sim U(a,b)$.

Total area = 1

Length = $b - a$

Height = $1/(b - a)$



anything out is 0

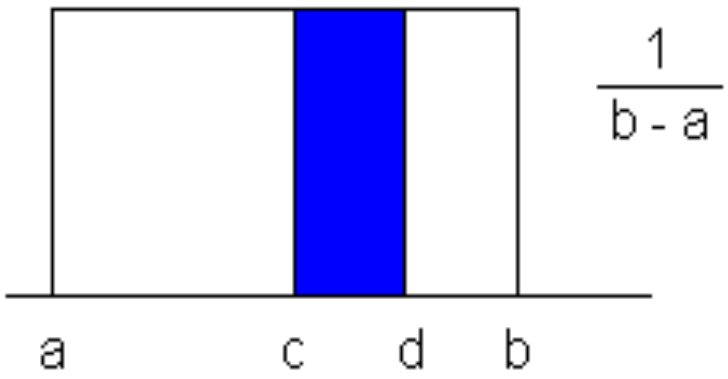
Prob. Distribution of Continuous r.v.

Thus the pdf is $f(X) = \frac{1}{b-a}$ for $a \le X \le b$

= 0, otherwise

For any values *c* and *d* between *a* and *b*

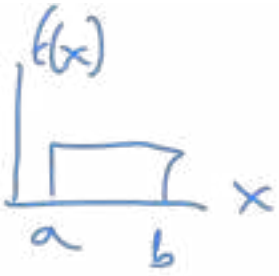
$$P(c < X < d) = \frac{1}{b - a}(d - c)$$



Exercise: Obtain the CDF.

Obtain the expressions for E[X] and Var[X] of r.v. X~U(a,b).

$$= \int_{-\infty}^{\infty} x \cdot f(x)\, dx \quad \text{(by definition)}$$



$$E[X] = \int_a^b \frac{x}{b-a} dx = \frac{1}{b-a}\left[\frac{x^2}{2}\right]_a^b = \frac{a+b}{2}$$

$$E[X^2] = \int_{-\infty}^{\infty} x^2 f(x) = \frac{1}{b-a}\int_a^b x^2 \cdot dx = \frac{a^2+ab+b^2}{3}$$

$$Var[X] = E[X^2] - E[X]^2 = \frac{a^2+ab+b^2}{3} - \frac{a^2+2ab+b^2}{4}$$

$$= \frac{a^2-2ab+b^2}{12} = \frac{(a-b)^2}{12}$$

Prob. Distribution of Continuous r.v.

Exponential Distribution

If the number of arrivals during an interval is Poisson distributed, then the interarrival times x are exponentially distributed. The pdf of x is:

$$f(x) = \lambda e^{-\lambda x}$$

where λ is the mean arrival rate and the mean interarrival time is $1/\lambda$.

Notation: $X \sim \text{Exp}(\lambda)$

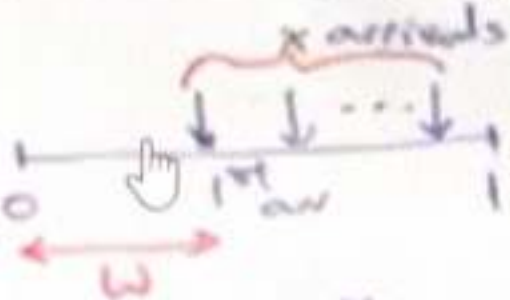
$$\mu = \frac{1}{\lambda}$$

If the number of arrivals in an interval is Poisson distributed, we show that the interarrival times are exponentially distributed.

If the number of arrivals in an interval is Poisson distributed, show that the interarrival times are exponentially distributed.

let X = no. of arrivals in 1 unit interval

$X \sim \text{Pois}(\mu)$, i.e. $P(X=x) = \frac{e^{-\mu}\mu^x}{x!}$



W is the 1st arrival time from $t=0$

∴ W = interarrival time and we want to find the p.d.f. of W

CDF $P(W \le w) = 1 - P(W > w)$

$= 1 - P(\text{no arrival in } [0, w])$

$= 1 - P(x = 0 \text{ in } (0, w])$

CDF $P(W \le w) = 1 - P(W > w)$

$= 1 - P(\text{no arrival in } (0, w])$

$= 1 - P(X = 0 \text{ in } (0, w])$

Poisson with mean μw

$= 1 - e^{-\mu w}$

$\therefore \text{pdf} = \frac{d}{dw} \text{CDF} = \boxed{\mu e^{-\mu w}}$

exponential dist

where μ = mean arrival rate

or $\frac{1}{\mu}$ = mean inter-arrival time

Prob. Distribution of Continuous r.v.

Exponential Distribution – examples:

Pdf:

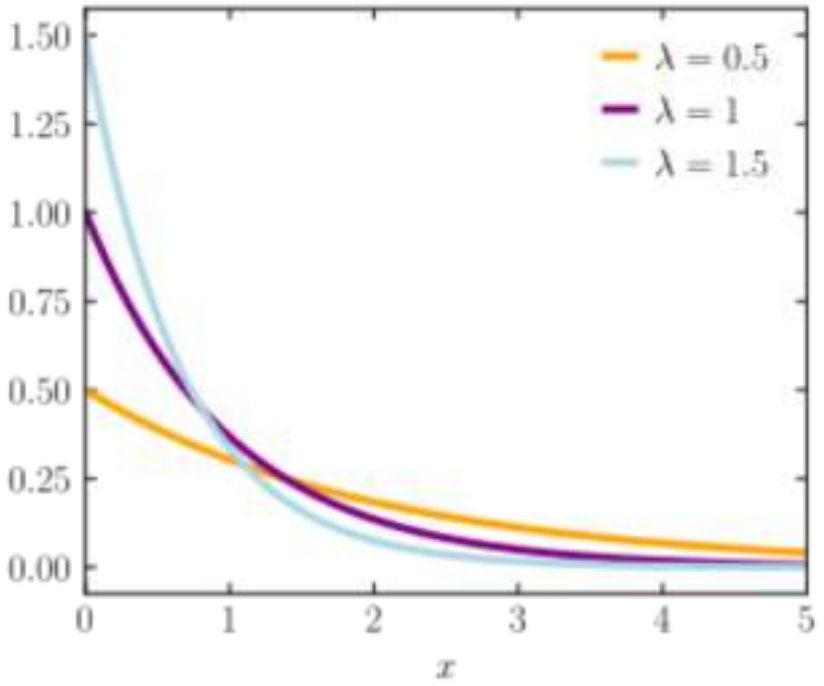
$$f(x) = \lambda e^{-\lambda x}$$

Expected value:

$$E[X] = \frac{1}{\lambda}$$

Variance:

$$\mathrm{Var}[X] = \frac{1}{\lambda^2}$$



Eg: Given that an average of number of customers walk into a shop per hour is 20 and the time between arrivals is exponentially distributed. After a customer arrives, what is the probability that the next customer will arrive within 1 minute?

avg arrived rate λ = 20/hr or $\frac{20}{60}$ /min

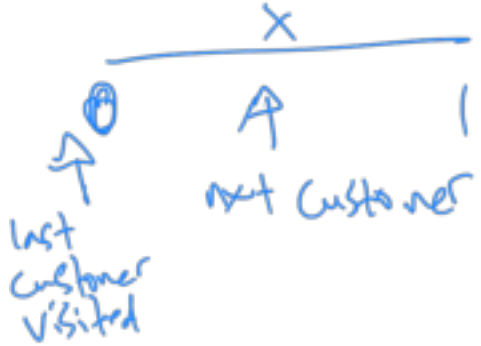
$\lambda = \frac{1}{3}$

$X \sim Exp(\lambda)$

$P(X < 1) = \int_0^1 \lambda e^{-\lambda x} dx$

$= [-e^{-\lambda x}]_0^1$

$= 0.283$



Prob. Distribution of Continuous r.v.

Normal Distribution

A common function for a continuous distribution is the normal (bell-shaped) curve. Its pdf depends on μ and σ, and is given by:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad \text{where } -\infty < x < \infty$$

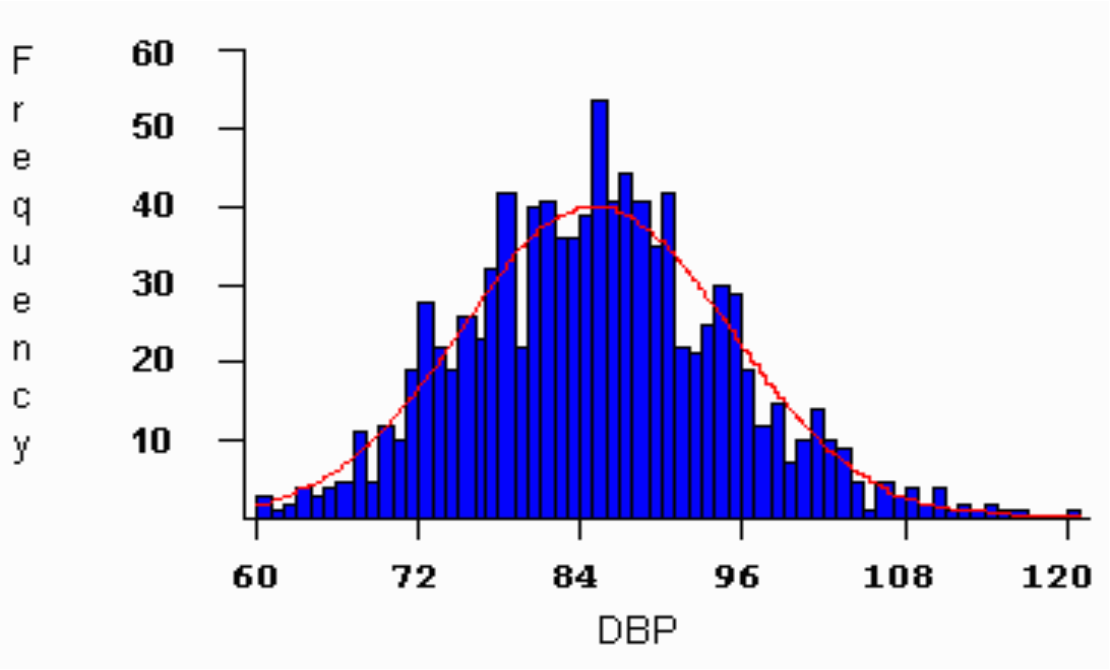
dun rid remember this formula

The normal distribution is by convention written as $X \sim N(\mu, \sigma^2)$. Eg: if $X \sim N(10, 16)$ then this implies that $\mu = 10$ and $\sigma^2 = 16$, and (hence $\sigma = 4$).

Prob. Distribution of Continuous r.v.

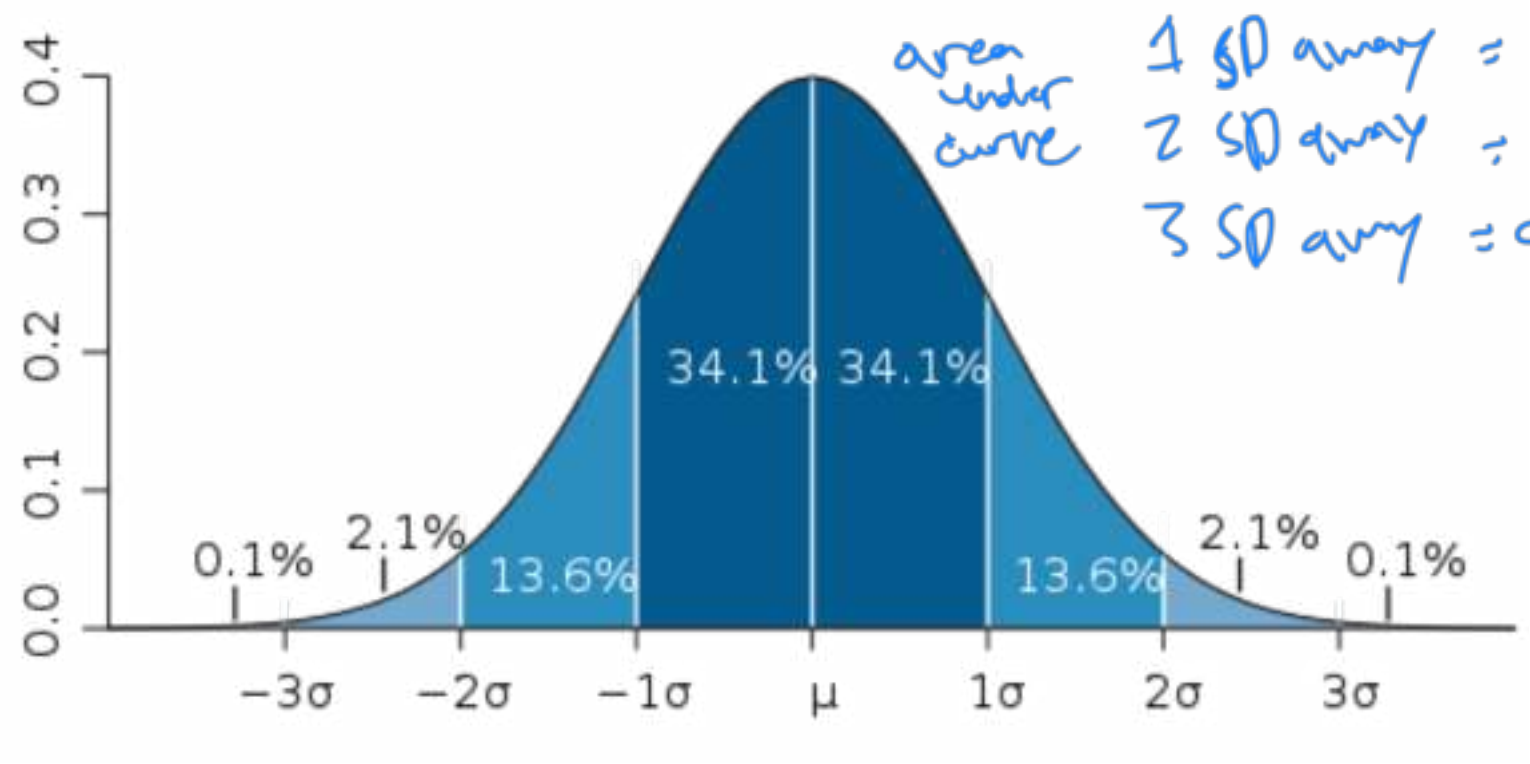
Eg: The distribution of blood pressure in patients can be approximated as a normal distribution with mean 85 mmHg and standard deviation 20 mmHg.

A histogram of 1,000 observations and the normal curve is shown below.



Prob. Distribution of Continuous r.v.

The normal distribution has about 68% of the observations lying within one standard deviation of the mean, 95% within two standard deviations and 99.7% within 3 standard deviations.



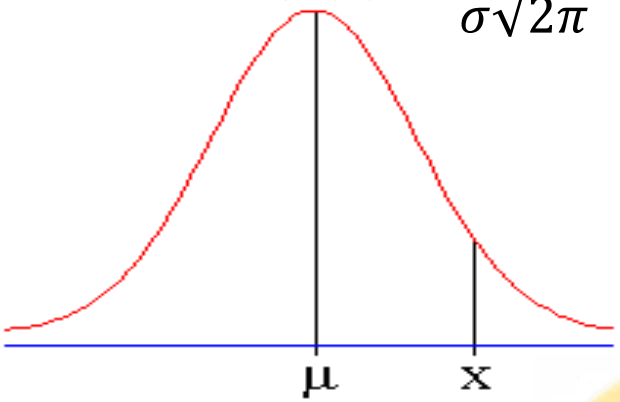
Prob. Distribution of Continuous r.v.

Standard Normal Distribution

To obtain the area under the density curve for a normal distribution, it is necessary to express any value of X in terms of the number of standard deviation units away from μ , i.e.

$$X = \mu + Z\sigma.$$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



This can be rearranged to give $Z = \dfrac{X - \mu}{\sigma}$

Prob. Distribution of Continuous r.v.

Since X is a r.v., so is Z. Using the formulas for functions of random variables we can obtain

$$E[Z] = \frac{1}{\sigma}E[X] - \frac{1}{\sigma}\mu = \frac{1}{\sigma}\mu - \frac{1}{\sigma}\mu = 0$$

$$\text{var}[Z] = \frac{1}{\sigma^2}\text{var}[X] = \frac{\sigma^2}{\sigma^2} = 1$$

Also it can be shown that Z has a normal distribution. Thus $Z \sim N(0, 1)$.

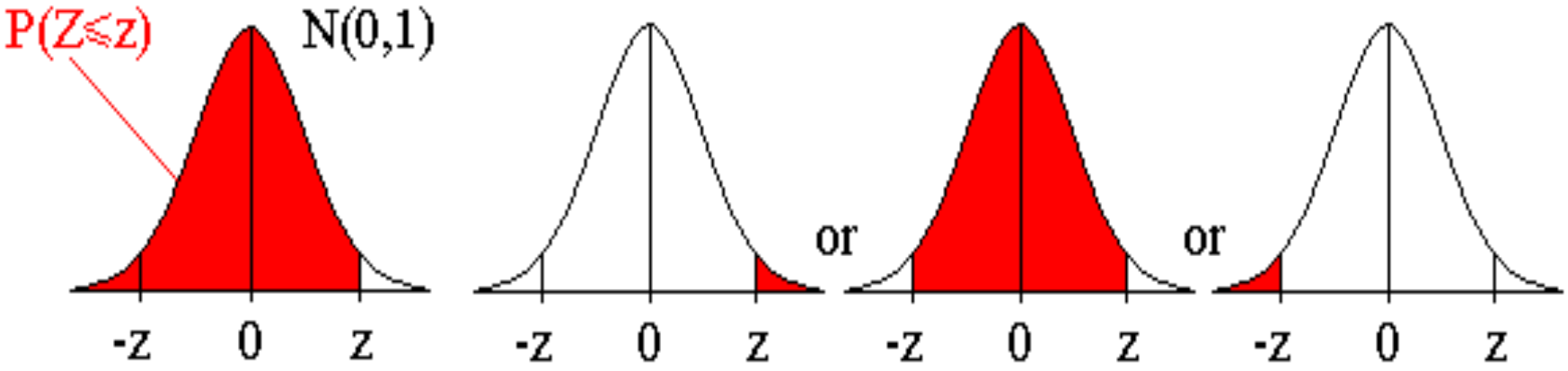
This is called the standard normal distribution.

↳ mean = 0
Var = 1

Prob. Distribution of Continuous r.v.

Most statistics textbooks include a table of the standard Normal distribution. It tabulates the area under the curve of the $N(0,1)$ density function.

Other normal tables may give other areas, e.g.



P(Z<0.21) = 0.5832

Table of Normal Distribution, $P(Z \le z)$

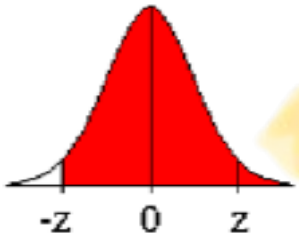
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7703	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Prob. Distribution of Continuous r.v.

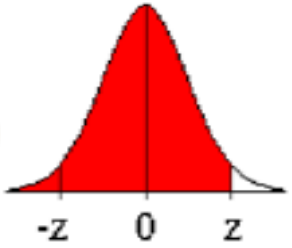
The $P(Z \leq z)$ table gives area under the curve from $-\infty$ to z, where $0 < z < 3.09$

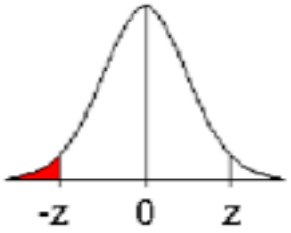
For other values, we perform some manipulations.

Eg:

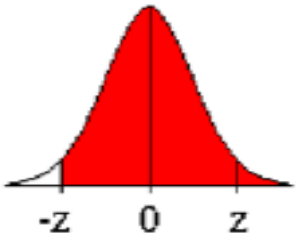


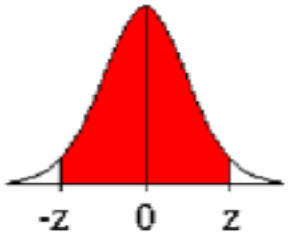
$$P(Z \geq -z) = P(Z \leq z)$$





$$P(Z \leq -z) = 1 - P(Z \geq -z)$$





$$P(-z \leq Z \leq z) = 1 - 2P(Z \leq -z)$$

Prob. Distribution of Continuous r.v.

Probabilities for normal distributions other than the standard normal distribution $N(0, 1)$ are obtained by using the formula:

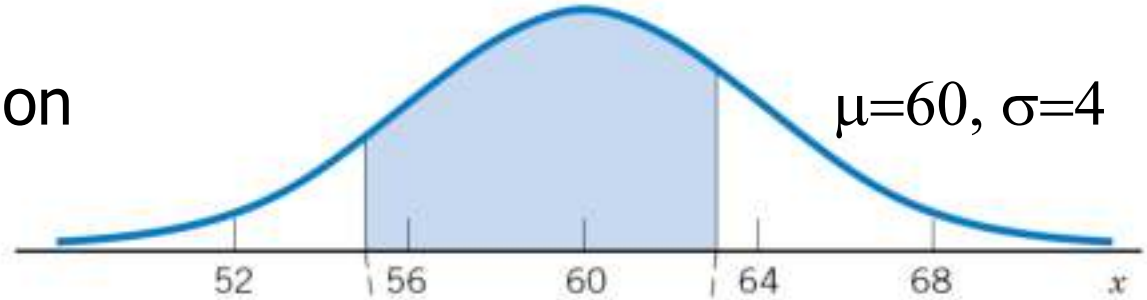
$$Z = \frac{X - \mu}{\sigma}$$

to convert from $X \sim N(\mu, \sigma^2)$ to $Z \sim N(0, 1)$ and then using the table of probabilities for the standard normal distribution $N(0, 1)$.

Prob. Distribution of Continuous r.v.

Normal distribution

$X \sim N(\mu, \sigma^2)$



$x = 63$

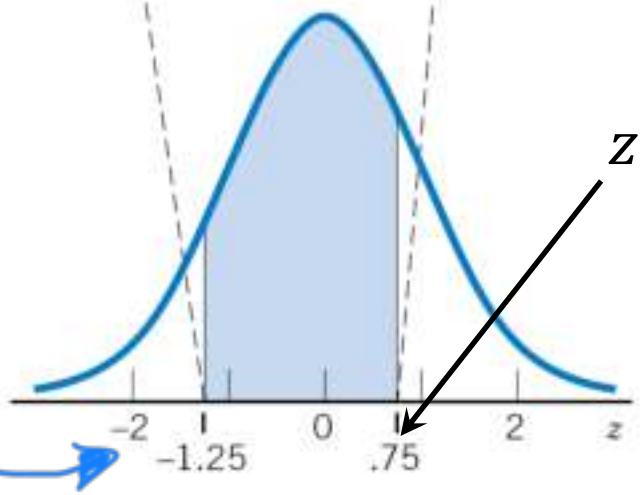
$z = \frac{55-60}{4}$

$= -\frac{5}{4}$

Standard Normal

$Z \sim N(0,1)$

$z = \frac{63-60}{4}$



where $Z=(X-\mu)/\sigma$

Prob. Distribution of Continuous r.v.

Eg: A machine fills bottles with 300 ml of soft drinks. However, the actual quantity filled varies according to the normal distribution with μ=298 ml and σ=3 ml.

What is the probability that an individual bottle contains less than 295 ml?

$X \sim N(298, 3^2)$

$P(X < 295) = P(Z < -1)$ ← See graph

$= 1 - P(Z < 1)$

$= 1 - 0.8413$

$= 0.1587 \approx 16\%$

$Z = \frac{295 - 298}{3}$

$= -1.00$

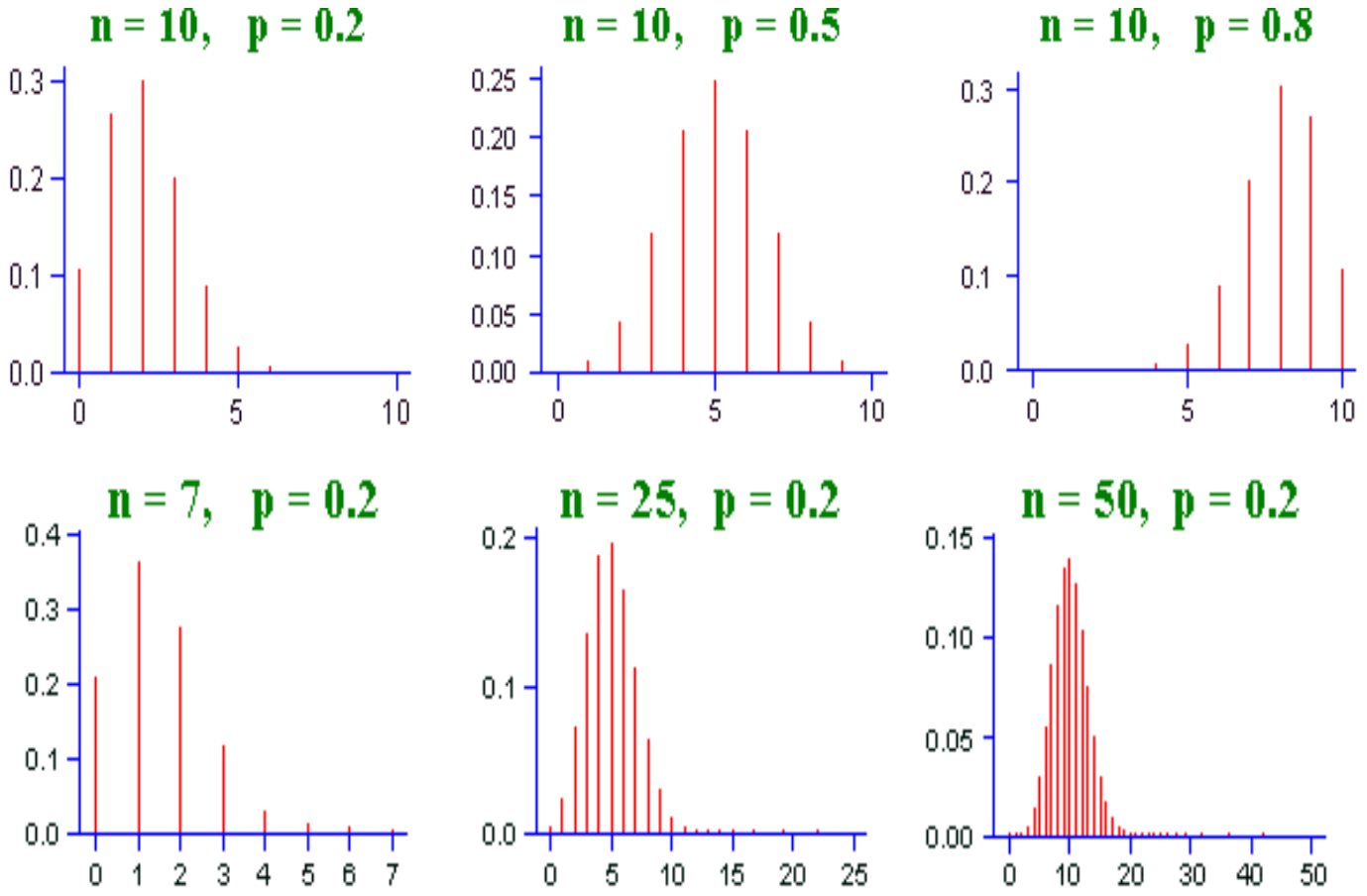
Prob. Distribution of Continuous r.v.

Solution

Let X denote the quantity in an individual bottle.
Given $X \sim N(298, 3^2)$, we want to find $P(X < 295)$

Approx. to Binomial Distribution

For large n or p near 0.5 the binomial distribution approximately follows the normal distribution.



Approx. to Binomial Distribution

If $X \sim \mathrm{B}(n, p)$ where n is large and p not too near 0 or 1, then X can be approximated by a normal distribution with $E(X) = np$ and $\mathrm{var}(X) = npq$, where $q = 1 - p$.

So $Z = \dfrac{X - np}{\sqrt{npq}}$ is approximately $N(0, 1)$.

This approximation is reasonably good when $np > 5$ and $n(1 - p) > 5$.

Approx. to Binomial Distribution

Continuity Correction and Accuracy

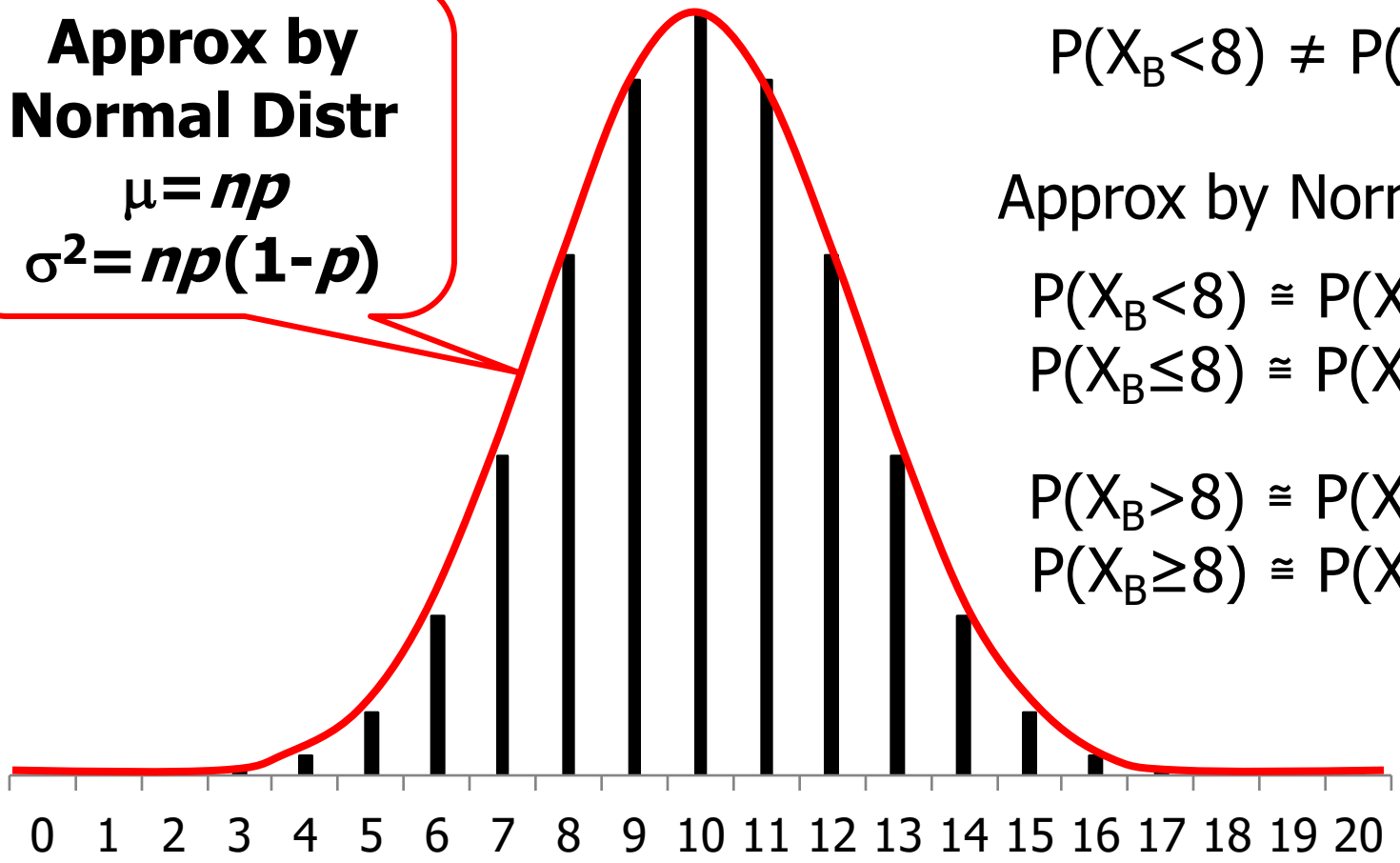
For more accurate values of binomial probabilities, the approximation is improved using the continuity correction. This method considers that each whole number occupies the interval from 0.5 below to 0.5 above it.

When an outcome X is to be included in the probability calculation, the normal approximation uses either $(X-0.5)$ or $(X+0.5)$.

Approx. to Binomial Distribution

Binomial Distribution $X_B \sim B(n,p)$

Approx by Normal Distr
$\mu = np$
$\sigma^2 = np(1-p)$



Binomial r.v.

$P(X_B<8) \neq P(X_B \leq 8)$

Approx by Normal distr

$P(X_B<8) \cong P(X_N<7.5)$
$P(X_B \leq 8) \cong P(X_N<8.5)$

$P(X_B>8) \cong P(X_N>8.5)$
$P(X_B \geq 8) \cong P(X_N>7.5)$

Approx. to Binomial Distribution

Eg: In a particular faculty 60% of students are men and 40% are women. In a sample of 50 students what is the probability that more than half are women?

let X = number of women in the sample n assume $X \sim B(50, 0.4)$

$E(X) = np = 20$ $\quad np = 30$ $\quad Var(X) = npq = 12$

$50 \cdot 0.4 \quad 50 \cdot 0.6 \quad 50 \cdot 0.4 \cdot 0.6$

we approximate X by $X_N \sim N(20, 12)$

$P(X > 25) = P\left(Z > \frac{25-20}{\sqrt{12}} = 1.44\right)$

$= 1 - P(Z \leq 1.44)$

$= 1 - 0.9251 = 0.075$

$P(X > 25) = P(X = 26) + \ldots + P(X = 50)$

$\rightarrow 0.0573$

Approx. to Binomial Distribution

We need to find $P(X > 25)$. Note: not $P(X \geq 25)$

The exact ans. calculated from binomial probabilities is:

Approx. to Binomial Distribution

The approximate probability, using the continuity correction, is

$$P(X > 25) \approx P\left(Z > \frac{25.5 - 20}{\sqrt{12}}\right)$$
$$\approx P(Z > 1.59)$$
$$\approx 1 - P(Z \leq 1.59)$$
$$\approx 0.0559$$

more accurate approximation than the original approximation (without correction) of 0.075

Approx. to Binomial Distribution

The value 25.5 was chosen as the outcome 25 was not to be included but the outcomes 26, 27, …, 50 were to be included in the calculation.

Similarly, if the example required the probability that less than 18 students were women, the continuity correction would require the calculation

$$P(X < 18) = P\left(Z < \frac{17.5 - 20}{\sqrt{12}}\right)$$

Summary

- Random Variables
- Expected Values and Variance
- Discrete Distributions
  1) Binomial
  2) Poisson
  3) Geometric
- Continuous Distributions
  1) Uniform
  2) Exponential
  3) Normal and Standard Normal
- Normal Approximation to Binomial

Alternate approach to obtain $\binom{n}{x}$

R1 1 1

R2 1 2 1

R3 1 3 3 1

• 1 4 6 4 1

• 1 5 10 10 5 1

• 1 6 15 20 15 6 1

1 7 21 35 35 21 7 1

1 8 28 56 70 56 28 8 1

⋮